CSIAM Trans. Life Sci.
doi: 10.4208/csiam-ls.SO-2024-0002

Vol. 1, No. 1, pp. 22-44
March 2025

Identifying Rarely Mutated Cancer Genes by Heterogeneous Network Embedding

Yurun Lu[1,2], Songmao Zhang[1] and Yong Wang[1,2,3,*]

[1] *CEMS, NCMIS, HCMS, MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.*
[2] *School of Mathematics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100049, China.*
[3] *Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, 310024, China.*

Received 11 December 2023; Accepted 9 May 2024

Abstract. Cancer is a multifaceted disease caused by dynamic interaction between genetic mutations and environmental factors. Understanding the genetic mutations underlying the development and progression of cancer is the stepstone for developing effective treatments and therapies. However, these mutations occurred in only a small fraction of cancer patients and it is extremely difficult to associate with cancer. Here, we propose MutNet, a heterogeneous network embedding method which integrate biomolecular network with cancer genomics data. Using pan cancer genomic data from The Cancer Genome Atlas program and public protein-protein interaction and pathway data, MutNet identifies rarely mutated cancer genes often overlooked by conventional genetic studies. In addition, the unified vector representation of biological entities allows us to reveal the tumor type specific cancer genes, cancer gene modules, and potential relationships among different tumor types. Our heterogeneous network embedding method holds the promise for the underlying mechanisms of cancer and potential therapeutic targets.

AMS subject classifications: 92B20, 92-08, 68T07

Key words: Cancer genomics, cancer gene, network embedding.

1 Introduction

Cancer is a complex disease caused by a combination of genetic and environmental factors. The genetic alterations (mutation, amplification, deletion, etc.) can change gene's

[*] Corresponding author. *Email address:* ywang@amss.ac.cn (Y. Wang)

normal function, which in turn lead to the uncontrolled growth and division of cancer cells [3]. Identification of these cancer genes is a key goal of cancer genomic analysis and stepstone in the development of precision oncology and cancer therapeutics [15,33,37,47]. Some cancer genes are frequently altered across many different types of cancer and can be easily identified, such as TP53, KRAS, and BRAF [47]. They are well studied in cancer development and progression by disrupting cell cycle regulation, DNA repair, or signaling pathways [37] in biomolecular network. However, some mutated cancer genes are altered in only a small fraction of cancer patients or a particular cancer type or subtype and are often overlooked in genetic studies [15,47]. These rarely mutated genes may play a significant role in the development and progression of cancer through propagating information via biomolecular network. Therefore, it is in pressing need to develop novel methods to identify these rarely mutated genes.

Whole exome sequencing (WES) allows researchers to sequence all of the protein-coding regions of the genome, known as the exome, to identify genetic mutations associated with cancer [33]. With the rapid accumulation of whole exome sequencing data, several algorithms have been developed to detect genes that are significantly mutated in cancer, such as ActiveDriver [33], TUSON [9], MuSiC [10] and MutSigCV [25]. They used statistical framework to identify significantly mutated genes based on the frequency and distribution of somatic mutations across tumor samples. OncodriveFM [16], OncodriveFML [29], OncodriveCLUST [42], and OncodriveCLUSTL [1] calculate the likelihood that a given gene is under positive selection in tumor samples based on the frequency and distribution of somatic mutations as well as gene-gene interaction information. 20/20+ [45] uses a machine learning algorithm to analyze WES data and predict driver genes that are likely to contribute to cancer development.

Despite the success of the above computational methods, there is still room to borrow information from gene interactions to boost rarely mutated cancer gene discovery. It’s well known that genes play crucial roles in various biological processes by acting in concert with each other within signaling and regulatory pathways, as well as in protein complexes [22]. The coordinated action of multiple genes allows cells to carry out complex functions such as growth, differentiation, and response to environmental stimuli. Therefore, network-based methods, such as HotNet2 [26] and OMEN [46] have been developed to detect cancer genes based on the interaction network and mutation patterns observed in WES data. Recently, EMOGI [36] integrates genomic data with other multi-omics data with graph convolutional networks to identify cancer genes. P-NET exploits hierarchical structure of biological pathways with convolution neural network to reveal molecularly altered candidates [13].

However, the above network-based methods are limited to their capacity of single type of molecules and associations, such as protein-protein interactions (PPI) and gene co-expression relationships, and cannot capture the full complexity of biological systems operating at many different levels [4, 24, 52]. We note that gene functional databased including pathways [48], gene ontology [44] provide rich information from different aspects. One way to capture all the information is constructing a heterogeneous network to allow multiple types of molecules and relationships. Even sample's meta information can be incorporated as additional nodes such as tumor types. Integrating this heterogeneous network with genomic data may help identifying rarely mutated gene. Furthermore, representation learning in heterogeneous network allows high level representability and scalability for integrated analysis [19,30] and drives advancements in various fields, such as social network analysis, recommendation systems, and bioinformatics [5].

In this paper, we propose MutNet, a heterogeneous network representation-learning method to integrate pan cancer genomic data with biomolecular network to identify candidate cancer genes. We first constructed a heterogeneous network with five different types of nodes including tumor types, samples, mutations, genes and pathways. The network is then embedded into a latent space with the genomic features and functional features of nodes preserved with a semi-supervised representation method, metapath2vec [12]. After all the samples, mutations, genes, pathways together with tumor types are embedded into a latent vector space, MutNet allows many flexible downstream tasks by measuring vector similarity. MutNet prioritizes cancer genes from large pan-cancer genomic data from the Cancer Genome Atlas (TCGA) and outperforms eleven existing methods. In particular, MutNet predicts 57 novel rarely mutated cancer genes (11 were reported in at least one dataset). Those cancer genes interact or co-pathway with known highly mutated cancer genes, rather than have a high mutation rate themselves. Moreover, MutNet identifies eight cancer gene modules associated with different function and pathway, tumor type-specific cancer genes, and potential relationship among tumor types based on the unified vector representation.

2 Methods

2.1 Identification of cancer genes with MutNet

MutNet is based on network representation learning and takes advantage of genes' PPIs and pathways to identify candidate cancer genes. As shown in Fig. 1, samples' genomic features including their somatic mutations are first integrated with the PPI network and pathways to construct a heterogeneous network. MutNet then identifies candidate cancer genes in the heterogeneous network by propagating the relationships among the genomic features, PPIs, and pathways through semi-supervised representation learning. The details are described in the following sections.

2.1.1 Inputs

MutNet takes the whole exome sequencing (WES) data of samples, protein-protein interaction network, and the biological pathways as inputs.

WES is powerful for identifying genetic mutations associated with inherited diseases, such as cancer and widely used for genetic diagnosis, genetic counseling, and personalized medical treatment for patients with inherited diseases. Particularly, we used the

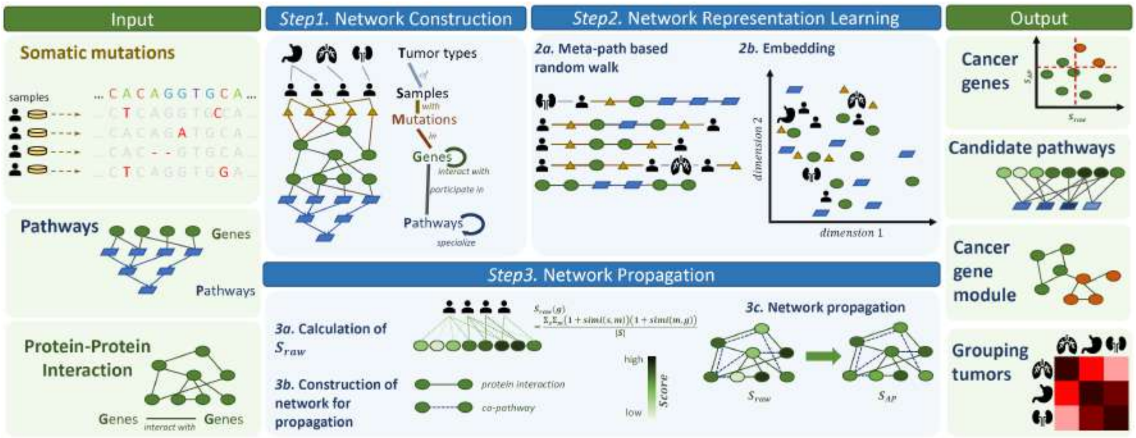

Figure 1: Schematic diagram of MutNet framework. MutNet takes whole exome sequence data from samples along with biological networks including PPIs and pathways as input. MutNet outputs candidate cancer genes or pathways. The three major steps include: 1. Heterogeneous network construction, 2. Representation learning, and 3. Network propagation. With both genomic and functional features integrated, MutNet embeds nodes including samples, tumor types, mutations, genes, and pathways into a latent space, and prioritizes pan-cancer or tumor-specific candidate cancer genes and pathways.

WES data from TCGA collected by Tokheim *et al.* [45]. WES data is preprocessed into a heterogeneous network

$$\mathcal{G}_{WES} = (\mathcal{V}_{WES} = \{S,T,M,G\}, \mathcal{E}_{WES} = \{E_{S-T}, E_{S-M}, E_{M-G}\}), \tag{2.1}$$

where $\mathcal{V}_{WES}$ includes four distinct node types including the samples S, their corresponding tumor type T, and the somatic mutations they possess M, as well as mutations to the genes G, and the edges $\mathcal{E}_{WES}$ abstracts the heterogeneous relationships between the nodes, including E_{S-T} for a tumor type T of a sample S, E_{S-M} for a sample S with a mutation M, E_{M-G} for a mutation M in a gene G.

A protein-protein interaction (PPI) network is a representation of the interactions among different proteins in a biological system. It can be abstracted as a weighted undirected graph, with nodes representing genes and edges representing their interactions. The STRING database is a publicly available resource that provides information on protein-protein interactions for a wide range of organisms [41]. The database is based on a combination of experimental data and predictions from computational methods, and it includes interactions from a variety of sources, such as protein-protein interactions, co-expression, co-occurrence, and experiments. STRING allows researchers to identify key proteins and pathways that are involved in the development and progression of cancer. Physical interactions among genes are used for constructing network in this work. The PPI network is prepossessed into a homogeneous network

$$\mathcal{G}_{PPI} = (\mathcal{V}_{PPI} = G, \mathcal{E}_{PPI} = E_{G-G}), \tag{2.2}$$

where G represents the genes and E_{G-G} represents the physical interaction between the genes.

REACTOME [14] is a publicly available database that provides information on biological pathways. It covers a wide range of organisms and pathways, including metabolic pathways, signaling pathways, and disease-related pathways. A gene-pathway network is constructed by processing the database download from REACTOME. Pathways are connected into a hierarchical structure with known child-parent relationships, and genes are connected to pathways by their annotation

$$\mathcal{G}_{Pathway}=\left(\mathcal{V}_{Pathway}=\{G,P\},\mathcal{E}_{Pathway}=\{E_{G-P},E_{P-P}\}\right), \tag{2.3}$$

where $\mathcal{V}_{Pathway}$ includes two distinct node types including the genes G, and the pathways P, and the edges $\mathcal{E}_{Pathway}$ abstracts the hierarchical relationships between pathways as E_{P-P} for a pathway P_1 is a child pathway of another more generalized pathway P_2, and E_{G-P} for a gene G participates in a pathway P.

2.1.2 Network construction

To integrate the WES data with knowledge graphs, MutNet interagates $\mathcal{G}_{WES}$ with $\mathcal{G}_{PPI}$ and $\mathcal{G}_{Pathway}$ to constructs a heterogeneous network with five types of nodes and six types of edges

$$\mathcal{G}_{MutNet}=\left(\mathcal{V}_{MutNet}=\{S,T,M,G,P\},\mathcal{E}_{MutNet}=\{E_{S-T},E_{S-M},E_{M-G},E_{G-G},E_{G-P},E_{P-P}\}\right). \tag{2.4}$$

This allows a comprehensive and integrative exploration of the molecular mechanisms underlying cancer and a deeper understanding of the interactions between different genomic features and candidate cancer genes.

2.1.3 Network representation learning

Within the heterogeneous network, we use representation learning to propagate information over the network structure and derive embedded vectors for the nodes.

Taking the constructed heterogeneous network as input, we convert the information contained in the network structure into an embedding vector for each node in a low-dimensional space, such that the vector of each node in this space forms a signature of the node which is useful for modeling hidden associations. Metapath2vec [12] provides a powerful method for heterogeneous network representation learning with meta-path-based random walks followed by a node co-occurrence based word2vec [27] embedding.

Meta-paths represent specific sequences of node types within a heterogeneous network, designed to capture intricate high-order relationships across different node types. To capture the complex interplay among nodes of various types and scales, we define six distinct meta-paths for conducting random walks.

1. MP_{TSMGP} $(T-S-M-G-P-P-P\cdots)$: traverses all tumor types and samples to collect basic mutation information.

2. MP_{GMST} $(G-M-S-T-S-M-G-M-S\cdots)$: co-mutation for samples with same tumor type or within the same gene is captured by this meta-path.

3. MP_{PGMST} $(P-G-M-S-T-S-M-G-P-G-M\cdots)$: co-pathway of genes and mutations occurring in two co-pathway genes are depicted.

4. MP_{GG} $(G-G-G\cdots)$: this meta-path mainly captures the interaction relationship among genes.

5. MP_{GGMST} $(G-G-M-S-T-S-M-G-G\cdots)$: similar with 3, and mutations occurring in two interacted genes are depicted.

6. MP_{GGPP} $(G-G-P-P-G-G\cdots)$: functional information including PPI network co-pathway are captured in this meta-path.

We traverse nodes on the constructed network and generate a corpus $\mathcal{C}$ where each sentence is a sequence of nodes. Formally, a sequence $C_{MP}(u_0)\in\mathcal{C}$ with a meta-path MP and a initial node u_0 is generated from a stochastic process M, where M_t represents the node visited by the random walk at step t, and $M_0=u_0$. The transition probability from node u to node v along the metapath MP at step t is defined as

$$P_{MP}(u,v,t)=\begin{cases}\dfrac{1}{|k\in MP^{(t+1)}|(u,k)\in\mathcal{E}|}, & u\in MP^{(t)},\quad v\in MP^{(t+1)},\quad (u,v)\in\mathcal{E},\\ 0 & \text{otherwise},\end{cases} \tag{2.5}$$

where $MP^{(t)}$ is the node type at step t defined in MP. $C_{MP}(u_0)$ starts from u_0 and proceeds by iteratively transitioning from the current node u_t to the next node u_{t+1} according to the transition probabilities in P_{MP}. All possible initial nodes in a predefined meta-path are traversed and the random walk process is repeated N times for each initial node and meta-path to generate a corpus $\mathcal{C}$.

Through random walking based on pre-defined meta-paths within the constructed heterogeneous network, we capture high-order neighborhoods that encompass the co-occurrence of mutations and the co-functioning of genes in a series of node sequences. These nodes are subsequently embedded into a latent space, where genes and mutations exhibiting co-function or co-occurrence within the heterogeneous network tend to exhibit higher cosine similarity. Nodes embeddings in a unified latent space $X\in\mathbb{R}^{|\mathcal{V}\times d|}$ are then obtained by minimizing the following loss function:

$$\mathcal{L}=\Sigma_{v\in\mathcal{V}}\Sigma_{t\in T_v}\Sigma_{u\in N(t,v,w)}\log p(u|v;\theta), \tag{2.6}$$

where T_v is the set of node types, $N(t,v,w)$ denotes the neighbour nodes from type t within distance w to node v in corpus $\mathcal{C}$, and

$$p(u|v;\theta)=\frac{e^{X_u\cdot X_v}}{\Sigma_{k\in\mathcal{V}}e^{X_k\cdot X_v}},$$

where X_v is the embedding vector for the node v, and θ is the mapping from the nodes to the latent space and is optimized during training. A relatively high cosine similarity between two embedded vectors suggests that the corresponding nodes are similar in terms of their connections and topological properties within the network. Specifically, the optimization of the objective function can be accomplished by a word2vec process which was used for capturing close words in sentences.

Notably, tumor type and sample nodes within our network introduce meta-information about cancer beyond the molecular level, allowing for information sharing across various cancer types. This unified vector representation of biological entities within the heterogeneous network serves as a foundational model for cancer genomics and has the potential for diverse applications.

2.1.4 Network propagation

With the heterogeneous network embedded into a latent space, MutNet assigns MutScore to genes and identifies cancer genes in three steps: Calculation of S_{raw}, network construction, and network propagation. A weighted mutation frequency for each sample and gene is first calculated

$$WMut(s,g)=\Sigma_{m\in M_s(s)\cap M_g(g)}\big(1+r(s,m)\big)\big(1+r(m,g)\big), \tag{2.7}$$

where $M_s(s)$ is the set of mutations in samples, $M_g(g)$ is the set of mutations occur in gene g, and $r(s,m)$ $(r(m,g))$ is the cosine similarity of s and m (m and g) in the embedded latent space.

Then S_{raw} is defined as follows:

$$S_{raw}(g)=\frac{\Sigma_{s\in S}WMut(s,g)}{|S|}, \tag{2.8}$$

where S is the set of samples and $|S|$ is the set size.

Overall, the S_{raw} prioritize genes with relatively higher mutation frequency. Rarely mutated genes interact or co-pathway with highly mutated genes can also be a cancer gene, and in order to figure out these genes, we construct a network based on the PPI network and pathways. The adjacency matrix of the PPI network is constructed as

$$A_{PPI}(g_i,g_j)=\begin{cases}1, & (g_i,g_j)\in\mathcal{E},\\ 0, & \text{otherwise}.\end{cases} \tag{2.9}$$

And the adjacency matrix of the pathways-based network is defined as

$$A_{pathway}(g_i,g_j)=\begin{cases}\Sigma_{p\in P(g_i)\cap P(g_j)}\dfrac{1}{|G(p)|}, & g_i\neq g_j,\\ 0, & \text{otherwise},\end{cases} \tag{2.10}$$

where $P(g)=\{p\in P|(g,p)\in\mathcal{E}\}$ is the set of pathways that gene g participate in, and $G(p)=\{g\in G|(g,p)\in\mathcal{E}\}$ is the set of genes in pathway p.

With A_{PPI} and $A_{pathway}$ defined, the network for score propagation is defined as

$$A_{NP} = qA_{PPI} + (1-q)A_{pathway}, \tag{2.11}$$

where $q \in (0,1)$ is a parameter for constructing the graph, and $q = 0.2$ in our results. The network A_{NP} combines the interaction and co-pathway relationships among genes. Network propagation is conducted on the network with S_{raw} as initial information

$$v^{(0)} = \big(S_{raw}(g_1),\cdots,S_{raw}(g_n)\big), \tag{2.12}$$

and the propagated results are defined as S_{AP} with S_{raw} propagated on the network

$$v^{(t+1)} = (1-k)v^{(t)}\overline{A_{NP}} + kv^{(0)}, \tag{2.13}$$

where $\overline{A_{NP}}$ is the normalized network propagation metric A_{NP} whose sum of each row equals to 1, and $k \in (0,1)$ is a parameter for propagation, and $k = 0.33$ in our results. Then the propagated score can be obtained by iteration

$$S_{AP} = \lim_{t\to\infty} v^{(t)}. \tag{2.14}$$

2.1.5 Outputs

MutNet outputs candidate genes based on the S_{raw} and S_{AP} calculated in the Eqs. (2.8) and (2.14). Genes with high S_{raw} have high mutation rates themselves, while S_{AP} depict the average mutation rate passing through interaction or co-pathway relationships. Candidate cancer genes are identified as genes with both relatively high S_{raw} and S_{AP}. We select a threshold for the scores according to their distribution

$$T_{PCG} = percentile\big(S_{raw}(g)\,|\,g \in G,\theta\big), \tag{2.15}$$

where θ can be selected according to requirement, and $\theta = 0.9$ in our results. Then the set of predicted cancer genes are defined as

$$PCG = \{g \in G \,|\, S_{raw}(g) > T_{PCG}, S_{AP}(g) > T_{PCG}\}. \tag{2.16}$$

These two restrictions ensure that genes with a relatively high mutation rate and frequently mutated interaction genes are accurately identified as cancer genes.

2.2 Evaluation of cancer gene prediction

Since the pan-cancer data offer an extensive molecular and genetic feature shared across various cancer types [28], we conducted our experiments based on a comprehensive pan-cancer dataset derived from TCGA collected by Tokheim *et al.* [45] to test whether MutNet can predict cancer genes accurately. Methods assign a p-value or score to each gene and the predicted gene set can be versatile from strict to soft according to the selection of the threshold. We calculate AUROC (area under the receiver operating characteristic curve) as a metric for evaluating the ranking of genes with standard cancer set. To calculate AUROC, the model's true positive rate (TPR) is plotted against its false positive rate (FPR) at various classification thresholds. The area under the resulting curve is then computed, which gives an indication of the model's ability to distinguish between positive and negative samples.

Standard known cancer genes are collected from the following four different datasets:

- KEGG cancer pathway: `https://www.genome.jp/entry/pathway+hsa05200` [23].
- MutPanning: `http://cancer-genes.org` [11].
- OncoKB: `https://www.oncokb.org/cancerGenes` [7].
- Cancer Gene Cosmic: `https://cancer.sanger.ac.uk/cosmic` [40].

Results from methods including 20/20+, ActiveDriver, OncodriveFM, TUSON, OncodriveFML, MuSiC, MutSigCV and OncodriveCLUST are collected from Tokheim *et al.* [45]. For methods which output q-values for genes, $-\log q$ is used for ranking genes and evaluation. As for 2020+ which outputs three p-values for each gene for tumor suppressor gene, oncogene, and driver gene respectively, combining p-values of tsg p-value, oncogene p-value, and driver p-value with Fisher's method is calculated and used for gene ranking. Genes with $q < 0.1$ are defined as predicted cancer genes identified by these methods. Results from network-based methods including HotNet2 and EMOGI are not trained on the same WES dataset and obtained by following the literature. We collected genes with top 10% score from HotNet2 and EMOGI as predicted cancer genes. p-values or scores for genes calculated by OMEN are not available, so we collected the top 80 genes from OMEN to compare the precision, recall, and f1-score.

S_{AP} of genes are used for ranking genes for our method, MutNet, in evaluation. As results in different methods various in the genes they cover, we extend the results to the union set provides by all methods, and the absent genes in different methods are added to the tail of the ranked gene list.

2.3 Contribution of PPI network and co-pathway network

To make clear that the benefit from genetic features of PPI network or co-pathway network, we calculated the respective contribution of PPI network and co-pathway network for each node. With the S_{raw} representing the contribution of genomic features, contributions of the two networks are defined as

$$C_{PPI}(g) = \Sigma_{g' \in G} \overline{A_{PPI}}(g,g') S_{raw}(g), \tag{2.17}$$

$$C_{pathway}(g) = \Sigma_{g' \in G} \overline{A_{pathway}}(g,g') S_{raw}(g), \tag{2.18}$$

where $\overline{A_{PPI}}$ and $\overline{A_{pathway}}$ represent normalized adjacency metric of PPI network A_{PPI} and co-pathway network $A_{pathway}$ respectively.

2.4 Tumor type specific cancer gene

The representation learning aspect of MutNet enables the derivation of tumor type-specific cancer genes, which is crucial for shedding light on the distinct molecular mechanisms underlying various cancer types. To identify tumor type specific cancer genes, we first calculated tumor type specific score S_T for each tumor type t and gene g

$$S_T(t,g)=\frac{\Sigma_{s\in S,T_s=t}WMut(s,g)}{|\{s\in S\,|\,T_s(s)=t\}|}, \tag{2.19}$$

where $T_s(s)$ denotes the tumor type of sample s.

Then tumor type specific cancer genes for tumor t are defined as

$$\begin{aligned}TSG(t)=\{g\in PCG\,|\,S_T(t,g)>T_{TSG},p_t(\{WMut(s,g)\,|\,T(s)=t\},&\\ \{WMut(s,g)\,|\,T(s)\neq t\})<\theta'\},&\end{aligned} \tag{2.20}$$

where $T_{TSG}=percentile(\{S_T(t,g)|t\in T,g\in G\},\theta),p_t(\cdot,\cdot)$ represents the p-value of t-test statistics, θ' can be selected according to requirement, and $\theta'=0.05$ in our results.

2.5 Candidate cancer pathways

MutNet uses the unified embedded vectors of the genes to rank the pathways in terms of their importance in cancer. The unified pathways representation with other nodes in the heterogeneous network provides the chance for identifying candidate cancer genes. For each pathway in REACTOME used for network construction, score to measure the potential cancer pathways S^P_{raw} is calculated as follows:

$$S^P_{raw}=\frac{\Sigma_{g\in G_P(p)S_{raw}(g)}(1+r(g,p))}{|G_P(p)|}, \tag{2.21}$$

and

$$S^P_{AP}=\frac{\Sigma_{g\in G_P(p)S_{AP}(g)}(1+r(g,p))}{|G_P(p)|}, \tag{2.22}$$

where $G_P(p)$ represents the set of genes participating in the pathway p annotated in REACTOME, $S_{raw}(g)$ and $S_{AP}(g)$ are calculated above, and $r(g,p)$ is the cosine similarity of gene g and pathway p in the embedded latent vector space.

Candidate cancer pathways set CCP is defined as

$$CCP=\{p\in P\,|\,S^P_{raw}(p)>T_{CCP},S^P_{AP}(p)>T_{CCP}\}, \tag{2.23}$$

where T_{CCP} is the threshold, and $T_{CCP}=0.1$ in our results.

2.6 Cancer gene modules

By representing tumor types and samples as embedded vectors in a unified latent vector space, MutNet is capable of uncovering potential relationships between different tumor types based on their genomic features, i.e. tumors are assigned with a vector representing their genetic features, and cosine similarity is used for association estimating. Cancer genes predicted previously are divided in to non-overlapping groups according to their embedded vectors in the unified latent space. Pair-wise genes similarity is first calculated and a mutual-kNN network is constructed by linking two genes if they are each other's KNNs, i.e. for any two genes g_i and g_j, if g_i is one of the k-nearest neighbors of g_j and g_j is one of the k-nearest neighbors of g_i, then an edge is added connecting g_i and g_j in the network. We select $k=15$ in our analysis, and use the Louvain algorithm in the Python network package `networkx` for community detection.

2.7 Genetic relationships between different tumor types

Genetic relationships between different tumors are inferred by combing tumor type specific S_T and the cosine similarity between the embedded vectors of tumor types and genes. The tumor type-specific genes' S_T is computed in Eq. (2.19), and genetic similarity between tumors is then computed based on the vectors which combine the genetic features of tumors and predicted cancer genes

$$D(t_i,t_j)=r\big(S_T(t_i,\cdot),S_T(t_j,\cdot)\big), \tag{2.24}$$

where $S_T(t,\cdot)\in R^{|PCG|}$ is the tumor type-specific score across all predicted cancer genes with tumor type t, and $r(\cdot,\cdot)$ is the cosine similarity of the two vectors.

3 Results

3.1 MutNet outperforms existing methods in identifying cancer genes

We trained MutNet on a high-confident WES dataset from Tokheim *et al.* [45] along with PPI network and gene-pathways network to identify cancer genes. We compared MutNet with eleven state-of-the-art methods for cancer gene prediction. These includes the methods ranking genes simply by their mutation rates and methods based on different types of machine learning algorithms, as well as statistical and bioinformatics approaches. Four of them, including ActiveDriver [33], MuSiC [10] and MutSigCV [25], and Mutation Rate, rank gene mainly according to their mutation rates. Five methods including OncodriveFM [16], OncodriveFML [29], OncodriveCLUST [42], TUSON [9], and 20/20+ [45] combine other information including mutation types, PPI, gene expression by traditional statistic models to optimize their prediction of cancer genes. Three network-based methods include HotNet2 [26], OMEN [46], and EMOGI [36]. EMOGI [36] is a supervised machine learning method taking multi-omics data including WGS, methylation, and RNAseq along with PPI network as input, and other methods takes genomic data such as WES or WGS along with open access database to rank genes.

We evaluated the performance of each method using a variety of metrics, including AUROC, recall, precision, and f1-score. To ensure that our validation is comprehensive, four different known cancer gene sets (KCG), including Cancer Gene Cosmic, MutPanning, OncoKB, and KEGG cancer gene sets, are collected as approximate gold standard positives for evaluation. Details for collecting of predictions from other methods and calculation for metrics are described in Section 2.2. As shown in Fig. 2A, MutNet overall outperforms mutation rate-based methods, multi-information-based methods, and network-based methods in predicting known cancer genes across the four gold standard datasets. Next, we used KEGG pathway as gold standard since it is known as a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for cancer. Taking advantage of knowledge integration MutNet obtains the highest AUROC in predicting functional cancer genes collected from KEGG pathway database [23] comparing with all other results. This demonstrates MutNet's ability to integrate multiple source information and boost the remote associations. EMOGI performs the best in AUROC for Cosmic and OncoKB cancer gene set due to the fact that it integrates paired multi-omics data including DNA methylation and gene expression by supervised learning from known cancer genes.

The comparison of AUROC among methods suggested that MutNet provides a high-quality ranking of candidate cancer genes. We then evaluated the Predicted Cancer Genes sets (PCG) generated from different methods for their recall, precision, and f1-score. In total 702 candidate cancer genes are selected by MutNet (Fig. 2B). 2020+ and TUSON provide smaller predicted cancer gene sets and reaches relatively high precision and f1-score, and MutNet outperforms other methods besides EMOGI in predicting cancer genes from KEGG cancer pathway, which suggests that MutNet is powerful in predicting rarely-mutated cancer genes (Fig. 2C and [55, Figs. S2x]). Moreover, across methods comparation shows that MutNet reaches the highest mean correlation with all other methods, which suggests MutNet's high consistency with other methods ([55, Fig. S2C]).

3.2 MutNet benefits from integrating genomic and functional features

MutNet involves in two main steps, the heterogeneous network representation learning for integrating genomic and functional features to obtain a foundation representation, and the network propagation for further enhance functional features for cancer gene prediction. Utilizing the foundational representation within the heterogeneous network, we calculate a weighted mutation rate (S_{raw}) that integrates both genomic and functional features (Eq. (2.8)). This foundational representation prioritizes mutations that are crucial to the samples, resulting in higher weights during the calculation of S_{raw} for those rarely-mutated but crucial genes. Genes ranked and selected based on S_{raw} demonstrate higher AUROC and f1-score compared to those selected solely based on mutation rate ([55, Fig. S2D]). These findings indicate that MutNet derives significant benefits from the heterogeneous learning approach and outperforms the mutation rate based methods.

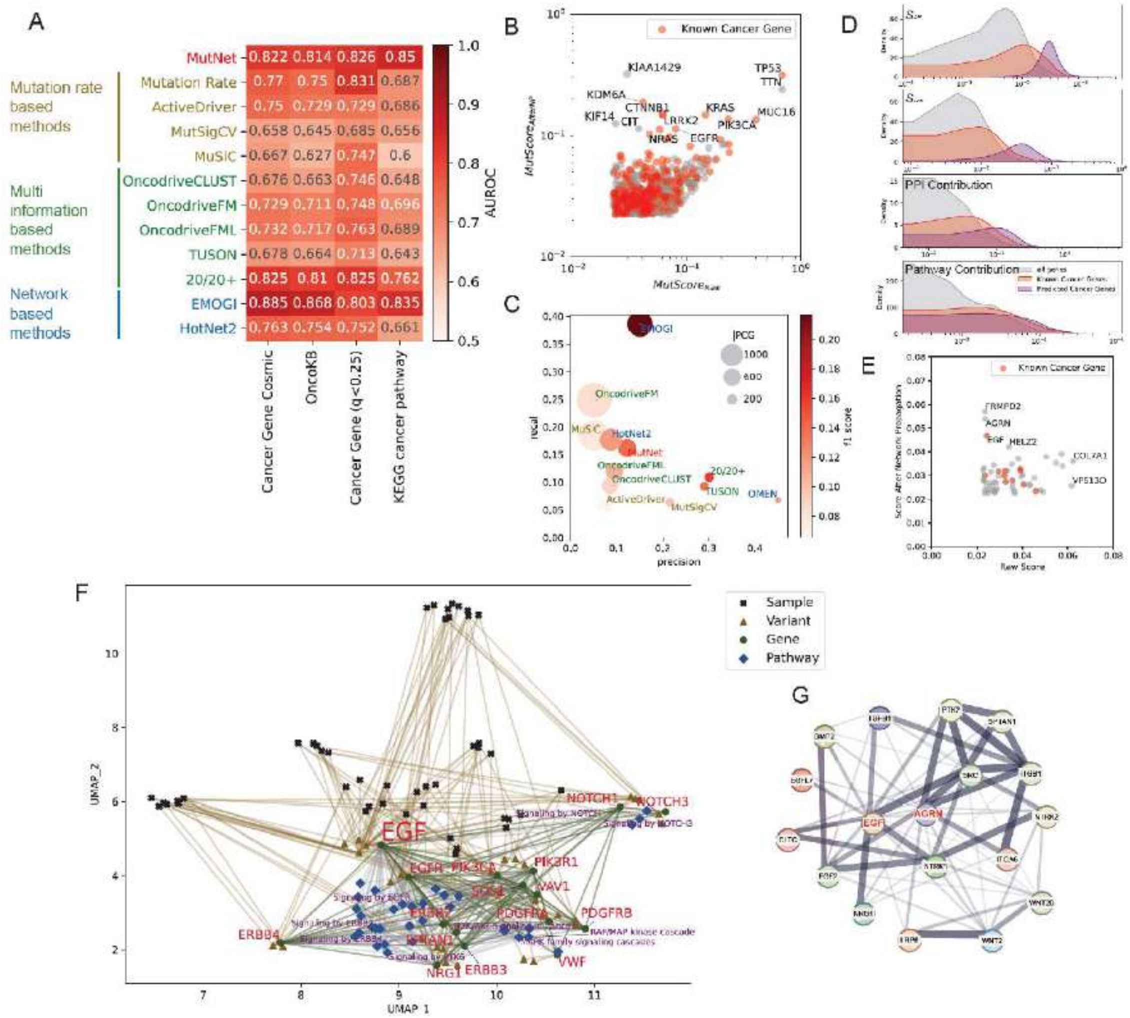

Figure 2: MutNet predicts functional cancer genes accurately. (A) AUROC of ranked cancer genes by different methods with four independent benchmark datasets. MutNet predicts functional cancer genes accurately. (B) The landscape of cancer genes predicted by MutNet. Overall, 702 genes are predicted as cancer genes by MutNet, of them 264 are recorded in at least one dataset. (C) Comparison of precision, recall, f1-score, and num of predicted cancer genes from different methods comparing with cancer gene from KEGG. MutNet provides high quality candidate cancer genes. (D) MutNet prioritizes genes with high PPI and high pathway contribution. Known cancer genes tend to have higher Sraw and network contribution. (E) 57 genes are uniquely predicted by MutNet, of them 11 are recorded in at least one dataset. (F) UMAP for the embedded vectors of an EGF-centric heterogeneous subgraph. Rarely mutated gene EGF interacts, co-mutates and co-pathways with other cancer genes, and is identified as a cancer gene in MutNet. (G) AGRN and EGF have physical interaction with 16 known cancer genes, and are predicted as cancer genes by MutNet.

HotNet2 [26] also involves in a network propagation on a PPI network to predict cancer genes. MutNet further integrates the PPI and co-pathway between genes to identify cancer genes. To make clear the contribution of genomic features (S_{raw}), PPI network, and co-pathway network in cancer genes identification during the network propagation, we deconvolute the network propagation output, S_{AP}, into three parts. We analyzed the differential distribution of the known cancer genes and the predicted cancer genes versus the all-gene background (Section 2.3). As shown in Fig. 2E, PPI network and co-pathways network helps to differentiate the cancer from the other genes. Moreover, by propagation on the network, rarely-mutated cancer genes are further selected out with relatively high S_{AP}. In summary, MutNet provides high quality candidate cancer genes by integrating the WES data with functional information including protein-protein interactions and gene pathways.

Furthermore, an ablation study is designed for the network propagation step in MutNet. The network is constructed from PPI network and genes co-pathway network, and the S_{AP} reaches a combination of genomic features and network structure. Adjusting the proportion of PPI network and genes co-pathway network illustrates the importance of the two gene functional networks, and the adjustment of the restart ratio in the network propagation reveals that both genomic features and the network structure contribute to the cancer genes identification. MutNet obtains the highest performance by combining different networks and genomic features ([55, Fig. S2E]). This suggests that all of the components are indispensable for an accurate identification of cancer genes.

3.3 MutNet identifies rarely mutated cancer genes

Overall, MutNet predict 702 candidate cancer genes (Fig. 2C, [55, Table S1]), and 264 of them are reported in at least one KCG set. 57 cancer genes are uniquely predicted by MutNet and are missed by other methods (Fig. 2E), and 11 of them were reported as a cancer gene in at least one of the KCG set. These genes are rarely mutated in cancers, and are ignored by other methods and the datasets. For example, EGF is involved in cancer pathways according to KEGG and has been uniquely identified as a cancer gene by MutNet. To investigate EGF's relevance to cancer, a subgraph was sampled from the heterogeneous network, consisting of samples, mutations, genes, and pathways closely related to EGF. As depicted in Fig. 2F, UMAP was utilized to display the embedded vectors of the nodes. EGF is found to be connected with several critical signaling pathways, including NOTCH signaling, EGFR signaling, and PI3K/AKT signaling in cancer. Furthermore, several cancer genes, such as ERBB4, EGFR, and NOTCH1, were found to interact and mutate in the same samples with EGF. Together this shows heterogeneous network allows MutNet to identify EGF as a rarely mutated cancer gene and holds the promise for better biological interpretation.

We showed that all the newly predicted genes have direct or indirect associations with cancer by literature search, and researches reveal their potential roles in cancer development ([55, Table S2]). For example, some studies have suggested that FRMRP2 may be a potential target for cancer therapy due to its role in regulating cell division [28]. Mutations in HELZ2 are associated with the prognosis in endometrial cancer [32]. The site and total immunohistochemistry score of COL7A1 expression in gastric cancer showed prognostic significance for OS and distant metastasis, respectively, which suggests that COL7A1 could be a novel biomarker with diagnostic and therapeutic value [35]. Prognostic analysis revealed high VPS13D expression to be associated with the adverse OS in acute myeloid leukemia [50]. As shown in Fig. 2G, the newly predicted cancer gene, AGRN, interacts with the uniquely predicted known cancer gene, EGF, and share 16 common cancer gene neighbors in PPI network [41]. Some studies also suggested that AGRN was significantly overexpressed in papillary thyroid cancer (PTC) and higher expression levels of AGRN were significantly associated with metastasis and poor prognosis of PTC patients [51].

3.4 MutNet identifies tumor type specific cancer genes

We utilized MutNet to identify cancer genes specific to different tumor types, leveraging the inclusion of tumor types as nodes in our heterogeneous network. With tumor types and samples, mutations, genes embedded into a single latent vector space, MutNet identifies 38 tumor type specific cancer genes for 24 cancers (Fig. 3).

MutNet identifies KRAS, the well-known pancreatic ductal adenocarcinoma (PDAC) driver gene, as the most PDAC-specific cancer gene as shown in our heatmap result (Fig. 3). KRAS is a commonly mutated gene in PDAC, a type of pancreatic cancer [38]. Mutations in KRAS are found in approximately 95% of PDAC cases, making it one of the most frequently mutated genes in this cancer type. KMT2C, TGFBR2, SMAD4 and

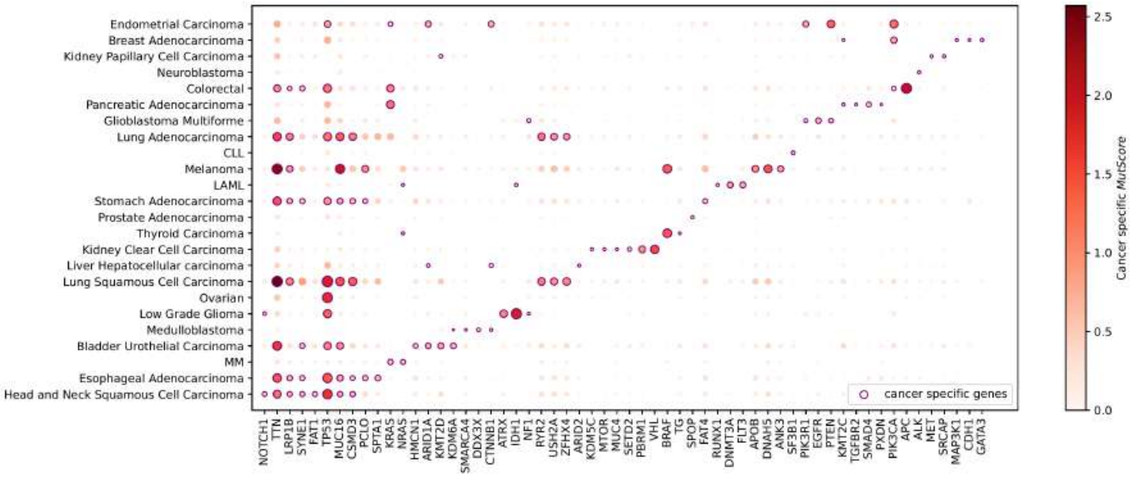

Figure 3: MutNet identifies tumor type specific driver genes. MutNet predicted cancer genes for each tumor type as well as tumor type specific cancer genes by taking pan-cancer whole exome sequence data along with samples' tumor types as input. Overall, 38 genes are identified as tumor type specific genes for 24 tumor types.

PXDN are also identified as PDAC-specific cancer genes. They have been implicated in the development and progression of PDAC through their roles in regulating cellular processes such as epigenetic modifications, signal transduction, and cellular differentiation [32, 53]. Other significant associations include IDH1 in low grade glioma, APC in colorectal cancer, VHL in kidney clear cell carcinoma, PTEN in endometrial carcinoma, DNAH5 in melanoma and so on.

3.5 MutNet reveals cancer associated pathways and modules

After single cancer genes are identified, we move on to their cooperation in latent modules. We identified 14 pathways that were considered as crucial in cancer development using the REACTOME database. Out of these 14 pathways, 7 are related to disease, and the others are involved in signal transduction processes, gene expression (transcription), or cell cycle (Fig. 4A, [55, Table S3]). On average, these pathways are annotated with

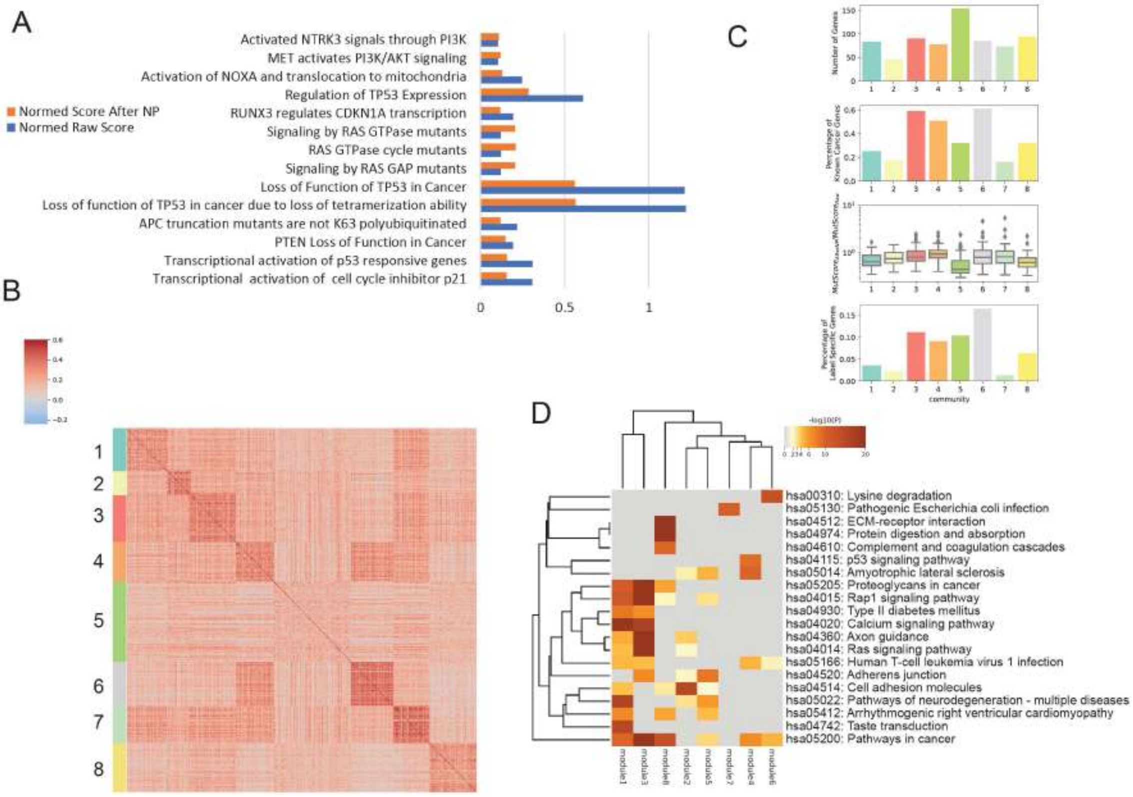

Figure 4: MutNet identifies candidate pathways and gene modules. A: MutNet prioritizes cancer related pathways based on genomic features. Highly mutated genes related pathways are identified as candidate pathways in cancer. B: Cancer genes are divided into 8 modules according to their embedded vectors in the latent space. C: Statistics and functional properties for the 8 modules. D: KEGG gene set enrichment analysis reveals the functional consensus and difference among the modules.

3.4 genes and contain 1.8 annotated cancer genes, highlighting their significance in the molecular mechanisms of cancer. This information can be used to better understand the functional relationships between genes and pathways, and to develop new therapeutic strategies for cancer treatment.

In addition to utilize pre-defined gene sets, MutNet also discovers new candidate cancer genes by creating a latent representation for all the genes in a unified space. MutNet allows a deeper understanding of the interactions between genes and reveals potential cancer gene modules. Cancer genes are connected by constructing a k nearest neighbor (kNN) network from their embedded vectors. Gene modules are defined by detecting communities in the cancer genes' network. Eight modules containing 702 predicted cancer genes are defined (Figs. 4B, 4C). This provides a comprehensive and integrative view of the cancer genes across different cancer types.

We then investigate the functional characteristics of the modules using KEGG pathway gene set enrichment analysis and revealed that those modules correspond to critical cancer hallmarks such as ECM-receptor interaction, axon guidance, cell adhesion molecules, and the Ras signaling pathway (Fig. 4D, [55, Table S3, Fig. S3]). Additionally, we highlighted the difference in functionality among the modules by the differentially enriched pathways. For example, pathways enriched in module 4 includes several pathways related to cancer and other diseases such as p53 signaling pathway, pathways in cancer, and human T-cell leukemia virus 1 infection. In addition, amyotrophic lateral sclerosis, a progressive neurodegenerative disease, is also among the identified pathways. These pathways may provide insights into the molecular mechanisms underlying cancer. Module 8 is enriched in various biological processes and signaling pathways that play important roles in cancer development and progression, including extracellular matrix (ECM)-receptor interaction, protein digestion and absorption, complement and coagulation cascades, and pathways in cancer. These processes and pathways are known to be associated with tumor invasion, angiogenesis, and immune response evasion, among others.

3.6 MutNet groups tumor types according to their vector representations in network

MutNet is capable of uncovering potential relationships between different tumor types based on their genomic features based on the unified latent embedding. This may help us to identify common patterns and shared characteristics across multiple tumor types, with the goal of improving our understanding of cancer as a whole [6]. Clustering samples based on their embedded vectors reveals that samples from the same cancer type exhibit high similarity. Interestingly, some patterns with relatively high similarity are observed across different cancer types, suggesting the possibility of a shared genomic similarity or relationship between different types of tumors (Fig. 5A). For example, neuroblastoma is closely associated with thyroid carcinoma as well as pancreatic adenocarcinoma.

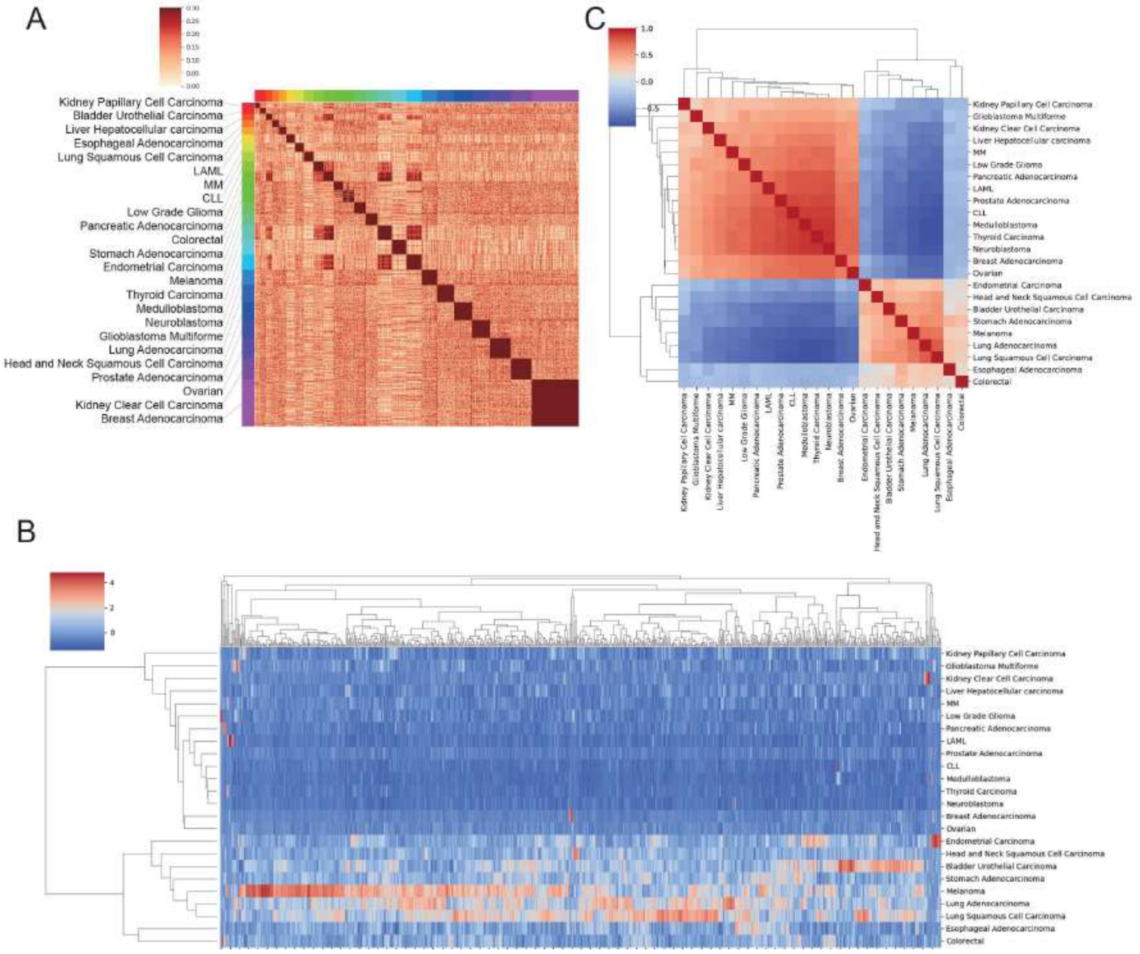

Figure 5: Unified embedded genomic features reveal the relationships among cancers. A: Samples with the same tumor types have high cosine similarity in latent space. Similarity across different tumor types can be revealed based on the embedded vectors. B: 24 types of tumors are clustered according to the genomic features of cancer genes in different tumor types. C: Tumor type specific S_T metric calculated for pan-cancer cluster analysis. Tumor types are divided into two clusters according to their overall mutation rate patterns.

Genetic similarity between tumors is then analyzed based on the unified vector representation and the predicted cancer gene set. Our results show that tumor types are divided into two clusters mainly according to their overall mutation rate (Figs. 5B, 5C). The overall mutation rate among tumor types can vary greatly depending on a variety of factors, such as the specific type of cancer, the stage of the cancer, the patient's age, and exposure to environmental carcinogens. The first cluster with overall higher tumor type-specific score, S_T, contains nine tumors including melanoma, lung adenocarcinoma, and lung squamous cell carcinoma, endometrial carcinoma, head and neck squamous cell carcinoma, bladder urothelial carcinoma, esophageal adenocarcinoma, colorectal, and stomach adenocarcinoma. Three most highly mutated tumors in the cluster, i.e. melanoma, lung adenocarcinoma, and lung squamous cell carcinoma, shared common genetic mutations including TP53, CDKN2A, BRAF, EGFR, and PIK3CA [17,31,43].

Cancers including pancreatic adenocarcinoma, LAML (acute myeloid leukemia), prostate adenocarcinoma, CLL (chronic lymphocytic leukemia), medulloblastoma, thyroid carcinoma, and neuroblastoma are all types of cancer that are generally considered to have a low mutation burden and are clustered in our analysis [20]. There are common features that can alter DNA methylation and gene expression, as exemplified by that pancreatic adenocarcinoma, LAML, and CLL are all associated with mutations in epigenetic regulators, such as DNMT3A, IDH1/2, and TET2 [8,18,39]. Mutations and activation of the hedgehog pathway have been found in medulloblastoma and thyroid carcinoma [34], and alterations in the RAS/RAF/MEK/ERK signaling pathway are commonly found in thyroid carcinoma and neuroblastoma [2]. These findings may have implications for understanding cancer biology and developing effective treatments that can target multiple tumor types.

4 Discussion

In this paper, we propose MutNet as a representation learning-based method for integrative analysis of genomic features with genes' functional information. One of the main advantages of MutNet lies in its ability to incorporate a wide range of genomic features, tumor types and sample annotations along with functional information. This allows for the integration of different types of data and knowledge, including somatic mutations, PPIs, and pathways, which can provide a comprehensive understanding of the molecular mechanisms driving cancer development and progression. As an example, we demonstrate that MutNet provides high quality candidate functional cancer genes and identifies rarely mutated cancer genes. In addition, by incorporating meta information from cancer patient samples, MutNet is able to reveal tumor type specific cancer pathways, potentially enabling the discovery of new targets for cancer treatment.

Moreover, MutNet's network representation learning strategy allows us to build a foundation model for many downstream tasks to identify complex biological interactions, providing a comprehensive view of cancer biology. By identifying potential cancer gene modules, MutNet can reveal how different genes work together to drive cancer development, potentially enabling the discovery of new biomarkers and therapeutics. Another key advantage of MutNet is its ability to associate different tumor types based on their genomic features. By identifying different genomic features that are unique to different types of cancer, MutNet can help researchers better understand the underlying biology of these diseases and develop personalized treatments and therapies. However, the computational demands of MutNet can be substantial, especially when handling extensive datasets. The construction, optimization, and training processes of the heterogeneous network involve a large number of nodes and edges, necessitating significant computational resources and time. Enhancements in efficient network representation learning theorems and advancements in hardware offer avenues to improve computational efficiency. Moreover, MutNet utilizes the shallow network representation learning method, metapath2vec, to integrate the heterogeneous data in the heterogeneous network. For the next step, we will further explore advanced deep-learning methods, such as HGT [21] and HAN [49], to extract more comprehensive information from the heterogeneous network.

MutNet is a highly adaptable framework that can effectively integrate various genomic features and identify significant biological signatures. In our current research, we have employed tumor types as sample's meta information, but the framework is flexible enough to accommodate other meta information such as pathological subtypes, survival time, drug information etc. Furthermore, MutNet supports meta information fusion analysis for individual samples, enhancing its versatility and applicability. Using a comparable architecture, we aim to expand the integration of diverse multi-omics data within the heterogeneous network. Our ongoing research involves adapting the MutNet framework to incorporate whole-genome sequencing data and explore the roles of noncoding drivers. We also aim to investigate the interactions between coding and noncoding drivers to gain deeper insights into cancer biology. Overall, MutNet's versatility and adaptability make it a valuable tool for exploring the complexities of genomics data and identifying significant biological features.

Acknowledgments

We acknowledge funding from the National Key Research and Development Program of China (Grant Nos. 2022YFA1004800, 2020YFA0712402), CAS Project for Young Scientists in Basic Research (Grant No. YSBR-077), and the National Natural Science Foundation of China (Grant No. 12025107).

References

[1] C. Arnedo-Pac, L. Mularoni, F. Muiños, A. Gonzalez-Perez, and N. Lopez-Bigas, *OncodriveCLUSTL: A sequence-based clustering method to identify cancer drivers*, Bioinformatics, 35:4788–4790, 2019.

[2] V. Asati, D. K. Mahapatra, and S. K. Bharti, *PI3K/Akt/mTOR and Ras/Raf/MEK/ERK signaling pathways inhibitors as anticancer agents: Structural and pharmacological perspectives*, Eur. J. Med. Chem., 109:314–341, 2016.

[3] A. Balmain, J. Gray, and B. Ponder, *The genetics and genomics of cancer*, Nat. Genet., 33(3):238–244, 2003.

[4] K. J. Bussey, L. H. Cisneros, C. H. Lineweaver, and P. C. W. Davies, *Ancestral gene regulatory networks drive cancer*, Proc. Natl. Acad. Sci. USA, 114:6160–6162, 2017.

[5] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins, *Next-generation machine learning for biological networks*, Cell, 173:1581–1592, 2018.

[6] Cancer Genome Atlas Research Network, *The Cancer Genome Atlas Pan-Cancer analysis project*, Nat. Genet., 45:1113–1120, 2013.

[7] D. Chakravarty et al., *OncoKB: A precision oncology knowledge base*, JCO Precis. Oncol., 1:1–16, 2017.

[8] S. M. Chan and R. Majeti, *Role of DNMT3A, TET2, and IDH1/2 mutations in pre-leukemic stem cells in acute myeloid leukemia*, Int. J. Hematol., 98:648–657, 2013.

[9] T. Davoli et al., *Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome*, Cell, 155:948–962, 2013.

[10] N. D. Dees et al., *MuSiC: Identifying mutational significance in cancer genomes*, Genome Res., 22:1589–1598, 2012.

[11] F. Dietlein et al., *Identification of cancer driver genes based on nucleotide context*, Nat. Genet., 52:208–218, 2020.

[12] Y. Dong, N. V. Chawla, and A. Swami, *metapath2vec: Scalable representation learning for heterogeneous networks*, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 135–144, 2017.

[13] H. A. Elmarakeby et al., *Biologically informed deep neural network for prostate cancer discovery*, Nature, 598:348–352, 2021.

[14] A. Fabregat et al., *The reactome pathway knowledgebase*, Nucleic Acids Res., 46:D649–D655, 2018.

[15] L. A. Garraway and E. S. Lander, *Lessons from the cancer genome*, Cell, 153:17–37, 2013.

[16] A. Gonzalez-Perez and N. Lopez-Bigas, *Functional impact bias reveals cancer drivers*, Nucleic Acids Res., 40:e169, 2012.

[17] H. Greulich, *The genomics of lung adenocarcinoma: Opportunities for targeted therapies*, Genes Cancer, 1:1200–1210, 2010.

[18] T. Hamidi, A. K. Singh, and T. Chen, *Genetic alterations of DNA methylation machinery in human diseases*, Epigenomics, 7:247–265, 2015.

[19] W. L. Hamilton, R. Ying, and J. Leskovec, *Representation learning on graphs: Methods and applications*, arXiv:1709.05584v3, 2018.

[20] D. Hao, L. Wang, and L. Di, *Distinct mutation accumulation rates among tissues determine the variation in cancer risk*, Sci. Rep., 6:19458, 2016.

[21] Z. Hu, Y. Dong, K. Wang, and Y. Sun, *Heterogeneous graph transformer*, in: Proceedings of the Web Conference 2020, ACM, 2704–2710, 2020.

[22] J. Iranzo, I. Martincorena, and E. V. Koonin, *Cancer-mutation network and the number and specificity of driver mutations*, Proc. Natl. Acad. Sci. USA, 115:E6010–E6019, 2018.

[23] M. Kanehisa and S. Goto, *KEGG: Kyoto encyclopedia of genes and genomes*, Nucleic Acids Res., 28:27–30, 2000.

[24] G. Kar, A. Gursoy, and O. Keskin, *Human cancer protein-protein interaction network: A structural perspective*, PLOS Comput. Biol., 5:e1000601, 2009.

[25] M. S. Lawrence et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes*, Nature, 499:214–218, 2013.

[26] M. D. M. Leiserson et al., *Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes*, Nat. Genet., 47:106–114, 2015.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, in: Advances in Neural Information Processing Systems, Curran Associates, Vol. 26, 2013.

[28] S. Moleirinho, A. Tilston-Lunel, L. Angus, F. Gunn-Moore, and P. A. Reynolds, *The expanding family of FERM proteins*, Biochem. J., 452:183–193, 2013.

[29] L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, and N. López-Bigas, *OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver muta-*

tions, Genome Biol., 17:128, 2016.

[30] J. Peng, G. Lu, and X. Shang, *A survey of network representation learning methods for link prediction in biological network*, Curr. Pharm. Des., 26:3076–3084, 2020.

[31] R. Rabbie, P. Ferguson, C. Molina-Aguilar, D. J. Adams, and C. D. Robles-Espinoza, *Melanoma subtypes: Genomic profiles, prognostic molecular markers and therapeutic possibilities*, J. Pathol., 247:539–551, 2019.

[32] B. J. Raphael et al., *Integrated genomic characterization of pancreatic ductal adenocarcinoma*, Cancer Cell, 32:185–203.e13, 2017.

[33] J. Reimand and G. D. Bader, *Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers*, Mol. Syst. Biol., 9:637, 2013.

[34] T. Rimkus, R. Carpenter, S. Qasem, M. Chan, and H.-W. Lo, *Targeting the sonic hedgehog signaling pathway: Review of smoothened and GLI inhibitors*, Cancers, 8:22, 2016.

[35] L. Roos et al., *Integrative DNA methylome analysis of pan-cancer biomarkers in cancer discordant monozygotic twin-pairs*, Clin. Epigenetics, 8:7, 2016.

[36] R. Schulte-Sasse, S. Budach, D. Hnisz, and A. Marsico, *Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms*, Nat. Mach. Intell., 3:513–526, 2021.

[37] C. J. Sherr, *Principles of tumor suppression*, Cell, 116:235–246, 2004.

[38] X. Shi et al., *Integrated profiling of human pancreatic cancer organoids reveals chromatin accessibility features associated with drug sensitivity*, Nat. Commun., 13:2169, 2022.

[39] B. Silverman and J. Shi, *Alterations of epigenetic regulators in pancreatic cancer and their clinical implications*, Int. J. Mol. Sci., 17:2138, 2016.

[40] Z. Sondka et al., *The COSMIC cancer gene census: Describing genetic dysfunction across all human cancers* Nat. Rev. Cancer, 18:696–705, 2018.

[41] D. Szklarczyk et al., *STRING v10: Protein–protein interaction networks, integrated over the tree of life*, Nucleic Acids Res., 43:D447–D452, 2015.

[42] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, *OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes*, Bioinformatics, 29:2238–2244, 2013.

[43] The Cancer Genome Atlas Research Network, *Comprehensive genomic characterization of squamous cell lung cancers*, Nature, 489:519–525, 2012.

[44] The Gene Ontology Consortium, *The gene ontology resource: 20 years and still GOing strong*, Nucleic Acids Res., 47:D330–D338, 2019.

[45] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, *Evaluating the evaluation of cancer driver genes*, Proc. Natl. Acad. Sci. USA, 113:14330–14335, 2016.

[46] D. Van Daele, B. Weytjens, L. De Raedt, and K. Marchal, *OMEN: Network-based driver gene identification using mutual exclusivity*, Bioinformatics, 38(12):3245–3251, 2022.

[47] B. Vogelstein et al., *Cancer genome landscapes*, Science, 339:1546–1558, 2013.

[48] B. Vogelstein and K. W. Kinzler, *Cancer genes and the pathways they control*, Nat. Med., 10:789–799, 2004.

[49] X. Wang et al., *Heterogeneous graph attention network*, in: The World Wide Web Conference, ACM, 2022–2032, 2019.

[50] J. Wei et al., *Identification the prognostic value of glutathione peroxidases expression levels in acute myeloid leukemia*, Ann. Transl. Med., 8:678–678, 2020.

[51] C.-C. Wu et al., *Integrated analysis of fine-needle-aspiration cystic fluid proteome, cancer cell secretome, and public transcriptome datasets for papillary thyroid cancer biomarker discovery*, Oncotarget, 9:12079–12100, 2018.

[52] Y. Yang et al., *Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types*, Nat. Commun., 5:3231, 2014.

[53] X. Zhou et al., *A systematic pan-cancer analysis of PXDN as a potential target for clinical diagnosis and treatment*, Front. Oncol., 12:952849, 2022.

[54] Y. Zhou et al., *Metascape provides a biologist-oriented resource for the analysis of systems-level datasets*, Nat. Commun., 10:1523, 2019.

[55] `https://github.com/YurunLu/MutNet`.