Mathematical Biosciences and Engineering

MBE, 20(12): 21643–21669.
DOI: 10.3934/mbe.2023958
Received: 27 August 2023
Revised: 19 November 2023
Accepted: 22 November 2023
Published: 06 December 2023

http://www.aimspress.com/journal/MBE

Review

Advances in computational methods for identifying cancer driver genes

Ying Wang[1], Bohao Zhou[1], Jidong Ru[2,*], Xianglian Meng[3], Yundong Wang[1] and Wenjie Liu[3,*]

[1] School of Computer Science and Engineering, Changshu Institute of Technology, Changshu 215500, China
[2] School of Textile Garment and Design, Changshu Institute of Technology, Changshu 215500, China
[3] School of Computer Information and Engineering, Changzhou Institute of Technology, Changzhou 213032, China

* **Correspondence:** Email: rujidong@126.com, liuwj@czust.edu.cn; Tel: 18814523450.

Abstract: Cancer driver genes (CDGs) are crucial in cancer prevention, diagnosis and treatment. This study employed computational methods for identifying CDGs, categorizing them into four groups. The major frameworks for each of these four categories were summarized. Additionally, we systematically gathered data from public databases and biological networks, and we elaborated on computational methods for identifying CDGs using the aforementioned databases. Further, we summarized the algorithms, mainly involving statistics and machine learning, used for identifying CDGs. Notably, the performances of nine typical identification methods for eight types of cancer were compared to analyze the applicability areas of these methods. Finally, we discussed the challenges and prospects associated with methods for identifying CDGs. The present study revealed that the network-based algorithms and machine learning-based methods demonstrated superior performance.

Keywords: cancer driver gene; CDG; computational methods; pathway

1. Introduction

Cancer is a genetic disease involving the malignant proliferation of cells caused by somatic mutations and clonal selection. Somatic mutations occur randomly throughout a person's life, and some specific genes mutate, providing selective growth advantages for normal epithelial cells and resulting in slow tumor growth. The number of polyclonal growing cells with other mutations continues to increase, eventually forming malignant tumors. Driver mutations provide a selective growth advantage for tumor proliferation and directly lead to cancer, while so-called "passenger

mutations" do not confer a direct effect in cancer cell proliferation, and the genes containing the driver mutations are "cancer driver genes (CDGs)" [1]. A typical cancer usually has 2–8 CDGs, as of September 2019, and 724 driver genes as recorded in the Cancer Gene Census (CGC) [2]. Therefore, CDG identification from a large number of genes in the cancer genome is an essential topic in cancer research to explore cancer pathogenesis.

Reviewing current methods can serve as a valuable guide for researchers, offering insights and experiences for peers dedicated to developing new methods. Therefore, we reviewed computational methods for identifying CDGs, aiming to more effectively analyze the characteristics and areas of excellence of various methods. To provide valuable insights to researchers, we categorized and summarized our findings systematically. We also traced the evolution of these methods and found that mutation frequency, biological networks, high coverage, high mutual exclusivity and machine learning algorithms play significant roles at various stages in developing CDG identification methods [3–5]. In addition, data are essential for all computational methods [6]. CDG identification methods are established based on different data sources, including somatic mutation data, gene expression profiles, biological networks stemming from genomics, transcriptomics and metabolomics. High-quality public databases help researchers uncover valuable insights. Different types of biological network models are constructed based on different relationships between molecules, thus representing different biological meanings [7]. Biological networks are usually combined with expression information to extract biological features between associated nodes, especially those related to CDGs. Moreover, it is essential to compare computational methods across different types of CDG identification methods. Such comparisons can further elaborate the performances of various method categories and offer valuable guidance to users of CDG identification methods.

The primary objective of this study was to review the research and progress in the field of CDGs based on computational methods. Additionally, we aimed to offer a comparative analysis of different CDG identification methods to benefit users. The study is organized as follows. Section 2 summarizes the classification of computational methods of CDG identification into four groups and the major frameworks of these four groups. Given the essential roles of data and algorithms in various computational methods, public databases and biological networks have been systematically compiled. Section 3 elaborates on data and associated computational methods for CDG identification. Section 4 analyzes the algorithms for CDG identification. Section 5 objectively analyzes the prediction fields of several methods by comparing and studying nine computational methods across eight different types of cancer. The last section discusses the opportunities and challenges of CDG identification. This manuscript serves as a valuable reference to researchers and guides those using these methods.

2. Classification of computational methods

In the field of CDG identification, mutation frequency is an intuitive feature that plays a significant role in identifying CDGs. Early-stage methods, based on statistical algorithms, were used to analyze somatic mutation frequency concerning the background mutation rate (BMR) [8–10]. However, these methods had limited effectiveness in detecting low-frequency mutations. Biological networks, which incorporate prior knowledge of the signaling pathway and can integrate gene expression data, proved instrumental in identifying CDGs with low-frequency mutations. Consequently, biological network-based methods were developed. Researchers studying the biological characteristics of cancer-driving genes observed that mutations in two driver genes could be

detrimental to cell proliferation or apoptosis, making them unsuitable for selection [11]. The methods were advantageous in terms of high coverage and high mutual exclusivity of CDGs developed and proved suitable for specific cancers [12]. Machine learning-based methods adopt more features and advanced algorithms to predict CDGs. They use typical biological characteristics of the genome, proteome, transcriptome and epigenome. Table 1 presents typical examples and classifications of CDG identification methods since 2010.

Table 1. Typical examples and classifications of CDG identification methods.

Example		Website
Gene mutation frequency-based methods	MutSigCV [8]	https://software.broadinstitute.org/cancer/cga/sites/default/files/data/tools/mutsig/MutSigCV_1.41.zip
	MuSiC [9]	https://github.com/ding-lab/MuSiC2
	OncodriveCLUST [10]	http://bg.upf.edu/oncodriveclust
	driverMAPS [13]	None
Network-based methods	DawnRank [11]	https://github.com/MartinFXP/DawnRank
	DriverNet [14]	None
	TieDIE [15]	None
	HotNet2 [16]	https://github.com/raphael-group/hotnet2
	MUFFINN [17]	http://www.inetbio.org/muffinn/
	MaxMIF [18]	https://sourceforge.net/projects/maxmif/files/
	VarWalker [19]	None
	DyTidriver [20]	https://github.com/weiba/DyTidriver
	OncoIMPACT [21]	http://sourceforge.net/projects/oncoimpact
Coverage and mutually exclusive feature-based methods	RME [22]	None
	CoMEt [23]	http://compbio.cs.brown.edu/software/comet
	Multi-Dendrix [24]	http://compbio.cs.brown.edu/software
	pathTiMEx [25]	https://github.com/cbg-ethz/pathTiMEx
	Dendrix [12]	http://cs.brown.edu/people/braphael/software.html
	TiMEx [26]	www.bsse.ethz.ch/cbg/software/TiMEx
	MEMo [27]	None
	nCOP [28]	https://github.com/Singh-Lab/nCOP
	EntroRank [29]	None
Machine learning methods	20/20+ [30]	https://github.com/KarchinLab/2020plus
	DriverML [31]	https://github.com/HelloYiHan/DriverML
	Agajanian etc. [3]	None
	Moonlight [32]	http://bioconductor.org/packages/MoonlightR/
	DeepDriver [33]	None
	GUST [34]	https://github.com/liliulab/gust
	DORGE [35]	https://github.com/biocq/DORGE
	cTaG [36]	https://github.com/RamanLab/cTaG
	CancerMine [37]	https://github.com/jakelever/cancermine
	LOTUS [38]	https://github.com/LOTUSproject/LOTUS

Note: The term "None" indicates that this method has no corresponding website.

2.1. Gene mutation frequency-based methods

Gene mutation frequency-based methods involve the analysis of somatic mutation frequency concerning the BMR by statistical analysis. These methods identify high-frequency mutations with significant differences as CDGs. Several methods have been proposed in this category. MuSiC selects high-frequency mutant driver genes that exhibit substantial increases over the BMR, classifying them as CDGs. MuSigCV [8] was the first algorithm to consider mutation rate heterogeneity. It can mutate more frequently than methods based on the inferred background mutation processes. Thus, it is more suitable for low-frequency driver gene identification because it reduces the identified false positives. ActiveDriver [39] is based on generalized linear regression to evaluate pairs of hypotheses for specific genes and their associated phosphosite regions. dNdScv investigates the synonymous substitutions originating from neutral evolution and calculates the nonsynonymous/synonymous variation rate (dN/dS) for each gene in the cancer. This method and covariates that affect the variation rate establish a null distribution for the expected number of nonsynonymous mutations. GISTIC2 identifies cancer-driven SCNVs (Signal Copy Number Variations) by estimating the background rate of somatic copy number variants. Then, it calculates the score for each region with the likelihood of reflecting the observed change frequency under the proposed background model. Genes with high scores are identified as driver SCNVs. Contrast rank is used to assess the overall risk score for adenocarcinoma based on genetic variants. DriverMAPS [13] stands out for its superior recognition performance and its ability to identify oncogenes (OGs) and tumor-suppressor genes (TSGs).

The accuracy of BMR estimation is the key to gene mutation frequency-based methods. The algorithm has low sensitivity and specificity due to the patient's uniqueness and the type of cancer. Hence, they are suitable for identifying high-frequency mutant CDGs. However, genome-wide surveys have revealed that most of the mutated genes fall into the low-frequency category (20–23%), limiting the performance of frequency-based methods.

2.2. Network-based methods

The network or pathway method assumes that the causal gene perturbation signal pathways drive the evolution of the cancer genomes. Cancer, being a complex disease, brings about many changes at the biological network level, allowing the identification of CDGs from the perspective of intergenic interactions. This method relies on prior knowledge of the signaling pathways, such as [40] VarWalker [19], HotNet2 [16], MUFFINN [17], MaxMIF [18] and others. VarWalker integrates the cancer genomic data and the protein-protein interaction (PPI) network using a random walk band, which restarts the algorithm in particular, adjusts gene length bias by resampling mutations in the individual genome and employs new network-based hierarchical methods stratifying the cancer subtypes [19]. HotNet2 adopts a random walk to spread gene mutation frequency across the PPI network. It identifies subnetworks with significant mutations and identifies potential driver genes based on mutation frequency and significance scores associated with the corresponding genes. MUFFINN proposes a pathway-centric mutation data analysis method, propagates gene mutation frequencies between direct neighbors on the PPI network and measures the impact of mutations in the adjacent genes using the sum of the maximum and full mutation frequencies of immediate neighbors [17]. MaxMIF ranks potential CDGs based on a gravity-based model where the mass and distance correspond to gene mutation scores and the relationship weights between the genes in the PPI network, respectively [18]. Shi et al. [41] assessed

and compared HotNet2, MaxMIF and MUFFINN using eight benchmark datasets, where HotNet2 showed the best overall performance.

As human interactome maps are constructed from mixed large-scale experimental data and are not specific to particular cell types, tissue types or conditions, they often suffer from incompleteness and errors. Thus, integrating an informative dataset at multiple omics levels (e.g., genome, transcriptome and epigenome) and developing an integration framework can provide a more comprehensive catalogue to prioritize driver genes at the network or pathway level.

The massive growth of cancer omics data has led to the emergence of a novel method that integrates genomic and transcriptomic data with biological networks. This approach focuses on identifying CDGs that affect downstream gene expression and interact to form functional modules. Therefore, this approach integrates abnormal gene expression and tissue-specific expression into biological network models and examines changes in gene expression to identify CDGs [42]. For example, DriverNet establishes a computational framework based on the impact of CDGs on mRNA expression levels, retains the mutated and differentially expressed gene nodes, constructs a bipartite map of the relationship between mutation and expression and uses a greedy algorithm to identify the differential gene combinations [14]. DawnRank trains a model using somatic mutation data, protein networks and gene expression profiles based on the random walk approach. It calculates the gene influence values based on the connectivity and the amount of differential expression of the downstream genes and ranks them at the individual patient level [11]. PARADIGM-SHIFT infers downstream pathways in cancer by integrating somatic mutations, CNVs and gene expression into an intact pathway using a belief propagation algorithm for breast cancer and pleomorphic glioblastoma [43]. TieDIE predicts the connectivity of the transcription factor target genes by integrating genomic and transcriptomic data into PPI networks and identifies the cancer-specific networks based on the existing literature [15]. DyTidriver identifies CDGs based on the node correlation and topology of the mutant network by introducing dysregulated gene expression, tissue-specific expression and variant frequency into the human functional interaction network [20]. Shikai et al. proposed a two-step method involving network diffusion and aggregation sorting. They combined the correlation of gene mutations, gene expression, the relationship between mutant genes and sample heterogeneous characteristics to construct a potential CDG sorting algorithm [44]. Suo et al. scored CDGs based on somatic mutations and the significance of gene differential expression. They selected high-frequency genes significantly differentially expressed from the neighboring nodes in the gene network [45]. Cd-CAP constructs subnetworks with conserved alteration patterns using a sample gene matrix based on the mutational information and gene-level data [46].

2.3. Coverage and mutually exclusive feature-based methods

These methods are based on mutual exclusivity and coverage of driver genes in signaling pathways, suggesting that a single driver gene in a pathway can promote cancer development and that two mutated genes are detrimental to cell proliferation or apoptosis and hence will not be selected [47]. These methods have advantages in terms of high coverage and mutual exclusivity of CDGs, ensuring that genes in a pathway cover as many patient samples as possible and that gene mutations in each pathway appear as distinct as possible in a single sample. For example, Dendrix employs the Markov Monte Carlo optimization method to suppress the overlap while improving driver gene coverage, ensuring exclusivity [12]. MEMO identifies a set of genes with a high frequency of mutations and

mutually exclusive properties [27]. Li et al. built functional networks of mutant genes and extracted the low-frequency mutant driver modules by integrating the functional similarity, coverage and mutual exclusion [48]. Gao Bo et al. proposed a *de novo* prediction method based on the exclusion, coverage and network topology for individual patients [49]. ModulOmics integrates PPIs, mutant interoperability, copy number alterations, transcriptional co-regulation and RNA co-expression into a single probabilistic model to identify cancer driver pathways or modules [50].

MEMO identifies only driver gene sets that are mutually exclusive and in the same pathway in the patient population. In contrast, the Dendrix algorithm identifies mutated subpathways that are mutually exclusive and have high coverage in the patient population. Most coverage and mutually exclusive feature-based methods are designed to identify driven pathways in a specific cancer type. However, HotNet2 is used for generalized cancer data and exhibits good recognition performance.

2.4. Machine learning methods

Machine learning methods perform excellently in many medical fields, such as coronary artery disease diagnosis [51]. Further, machine learning methods have become essential for predicting CDGs and mutation functions in modern biomedical research and have gained significant momentum in the last decade [52]. This category of methods typically involves extracting features from driver genes and training a classifier to predict these genes, such as [53,54]. These methods are trained using either pathogenic or neutral mutations.

Typical machine learning methods, including the support vector machine (SVM), random forest and Bayesian algorithm, have been widely used; and with the development of deep learning research, algorithms such as convolutional neural networks (CNNs) have also demonstrated strong performance in pattern recognition [55]. For example, the 20/20+ method selects features such as gene frequency, mutation type, expression level or replication time of genes in different cancer cells and predicts driver genes using a random forest approach [30]. DriverML uses the Rao score to calculate the impact of mutations on proteins, optimizes weight parameters and maximizes the score statistics of previously identified driver genes across pan-cancer training data [31]. Agajanian et al. [3] used 6389 validated cancer driver mutations and 12,941 passenger mutations. They utilized a 2570 mutation driver/passenger classification. For analysis, this study combined CNN-based learning features with embedding-based functional features and used a random forest approach for classification and identification purposes.

Machine learning methods, especially deep learning methods [56,57], have demonstrated excellent performance across various fields. For example, DeepDriver uses CNNs to extract information from mutation data and similarities, enhancing the driver gene prediction [33].

Some deep learning algorithms have also been applied to driver gene identification problems. Driver genes can be categorized as TSGs and OGs. OGs are activated by gain-of-function mutations, whereas TSGs are inactivated through loss-of-function mutations. Some machine learning methods are designed to identify OGs and TSGs. LOTUS identifies genes with high oncogenic potential by integrating various data types, including information about gene mutations and PPIs, using an SVM [38]. DORGE identifies TSGs and OGs based on the penalized regression method and 75 genetic and epigenetic features related to mutation, genomics, phenotype, epigenetics and their complements, as well as TUSON and CRISPR screening-only features [35]. GUST discovers OGs and TSGs with high accuracy (92%) based on the RF algorithm [34]. On the contrary, CancerMine presents a method for

gene-centric clustering of cancer types by weighting gene roles based on the number of supporting manuscripts and using a high-precision classifier [37].

In the aforementioned methods, GUST, DORGE, cTaG and CancerMine have been developed for identifying TSGs and OGs. In contrast, other methods are used to classify cancer driver mutations. In terms of performance, the accuracy of GUST is consistently higher than 20/20+ [34]. Additionally, three methods, including TUSON, MutSigCV and 20/20+, have demonstrated superior performance compared with the five other methods. These include ActiveDriver, MuSiC, OncodriveClust, OncodriveFM and OncodriveFML [17]. When comparing network-based algorithms (Moonlight, Ncop, OncoIMPACT, HotNet2, MaxMIF, MUFFINN and NetSig,), MutPanning and frequency-based algorithms (driverMAPS, WITER and DriverML), HotNet2 and driverMAPS demonstrated the best overall performance [42]. Machine learning methods can discover driver genes that are difficult to detect using methods based on gene mutation frequency. They can identify many driver genes with very low mutation rates.

Cancer genomics data include somatic mutations, transcriptomes, methylation and proteomics from patient tumors and their matched normal tissues. Some methods have been developed by integrating multiomics data, indicating an important research direction for the future [52,58–62].

The use of various omics data types in different methods highlights the significance of algorithms. Therefore, the data and algorithms are introduced in detail next.

3. Data for CDG identification methods

3.1. Public database

The highly credible data were downloaded from a public database and used more than twice in the methods discussed earlier. The databases used for CDG identification are listed in Table 2.

Table 2. Public databases, data types and websites used for CDG identification.

Database	Type	Website
TCGA [63]	Genomic variation, mRNA expression, miRNA expression and methylation	http://cancergenome.nih.gov
GEO [64]	Gene expression	https://www.ncbi.nlm.nih.gov/geo/
cBioPortal	Somatic mutations, DNA copy number alterations, mRNA and microRNA expression, DNA methylation, protein abundance and phosphoprotein abundance	http://www.cbioportal.org
COSMIC v98 [2]	Somatic mutations	http://cancer.sanger.ac.uk/cosmic
ICGC Release 28 [65]	Abnormal gene expression, somatic mutations, epigenetic modifications and clinical data	https://icgc.org

Continued on next page

Database	Type	Website
CCLE [66]	Gene expression, chromosomal copy number and massively parallel sequencing data	https://portals.broadinstitute.org/ccle/
DisGeNET v7.0 [67]	Gene-disease association and variation-disease association	http://www.disgenet.org/
NCG7.1 [68]	Cancer genes	http://ncg.kcl.ac.uk/
TARGET	Clinical annotation, gene expression, chromosome copy number analyses, epigenetics, miRNA profiling, whole-genome sequencing, whole-exome sequencing and mRNA-seq	https://ocg.cancer.gov/programs/target
Cancer3D v2 [69]	Somatic missense mutations	http://www.cancer3d.org/
dSysMap V2020_05 [70]	Gene mutations	http://dsysmap.irbbarcelona.org
ENCODE [71]	Chip-seq of transcription factors	https://www.encodeproject.org
NIH Epigenome Roadmap [72]	DNA accessibility, DNA methylation and RNA expression	http://www.roadmapepigenomics.org
FANTOM5 [73]	Transcripts, transcription factors and promoters	http://fantom.gsc.riken.jp/5/
GTEx [74]	Genotype tissue expression	http://www.gtexportal.org/

3.2. Biological networks

In recent years, many cancer studies have considered biological networks to interpret driver genes in cancer. Biological networks include protein interaction networks, gene transcription regulation networks and networks related to biological metabolism and signal transduction. These networks also include pathways involved in metabolism, gene expression regulation and signaling transduction. Therefore, resource reviews can help researchers choose the right network. Biological networks used in the typical CDG identification methods are listed in Table 3.

TCGA covers multiomics data from 33 types of cancer and over 11,000 patients with cancer [9]. GEO stores approximately 112,752 libraries generated by 19,692 laboratories, including 3,027,904 data samples from over 1600 biological sources [11]. cBioPortal provides multidimensional cancer genomics data for over 5000 tumor samples across 20 cancer studies [2]. COSMIC, combined with expert knowledge and a genome-wide database, is the largest and most comprehensive resource for somatic mutations associated with human cancer [75]. ICGC plays a significant role in cataloguing tumor genomic abnormalities across 50 different cancer types and/or subtypes, establishing itself as the largest public database of microarray data focusing on the unique genetic characteristics of individual tumor types [65]. The Cancer Cell Line Encyclopedia (CCLE) provides genomic data, analysis and visualization of 1457 cell lines [66]. DisGeNET contains 21,671 genes associated

with 30,170 diseases, features and clinical or abnormal phenotypes. It encompasses data on 369,554 variant diseases, including 194,515 variants associated with 14,155 diseases, features and phenotypes. The extensive dataset significantly contributes to the understanding of 1,134,942 genetic diseases [67]. NCG [68] contains information on replicability, evolution, PPI, miRNA-gene interaction, function and expression. This information has been extracted from 273 manually curated publications and covers data on 2372 cancer genes. Therapeutically Applicable Research to Generate Effective Treatments (TARGET) contains genomic, transcriptomic and epigenetic genomic data for the study of pediatric cancers. Cancer3D [69] contains 1,457,702 mutations in 9079 samples of 32 cancers mapped to 18,425 proteins. DSysMap [70] maps gene mutations associated with human diseases to protein structures and also includes interactions in the human interaction genome. ENCODE [71] systematically describes the transcription region and the associated transcription. The NIH Epigenome Roadmap [72] contains maps of DNA methylation in stem cells and primary isolated tissues, histone modifications, chromatin accessibility and small RNA transcripts, as well as a normal epigenome. FANTOM5 [73] is a complex multicellular biological database consisting of about 400 different cell types. It predominantly features primary mammalian cell types as well as a range of cancer cell lines. The dataset also encompasses sets of active transcripts, transcription factors, promoters and enhancers in tissues. GTEx [74] provides gene sequencing data from normal tissue. Pan-cancer multiomics resources are pivotal, providing abundant and comprehensive multidimensional data for the research and identification of CDGs.

Table 3. Biological networks used in CDG identification methods.

Biological network	Type	Website
HPRD Release 9 [76]	Protein interaction network	http://www.hprd.org
BioGRID 4.4 [77]	Protein, DNA and drug interaction network	http://thebiogrid.org
STRING12.0 [78]	Protein interaction network	http://string-db.org
iRefWeb [79]	Protein interaction network	ftp://ftp.no.embnet.org/irefindex/data
MINT [80]	Protein interaction network	http://mint.bio.uniroma2.it/mint/
IntAct 1.0.3 [81]	Protein interaction network	http://www.ebi.ac.uk/intact/
PINA [82]	Protein interaction network	http://cbg.garvan.unsw.edu.au/pina/
PhosphoSitePlusv6.7.1.1 [83]	Protein interaction network	http://www.phosphosite.org/
Phospho.ELM [84]	Protein interaction network	http://phospho.elm.eu.org
PTMcode 2 [85]	Protein interaction network	http://ptmcode.embl.de
Interactome3D 2020_05 [86]	Protein interaction network	http://interactome3d.irbbarcelona.org
3did [87]	Protein interaction network	https://3did.irbbarcelona.org/
Instruct [88]	Protein interaction network	http://instruct.yulab.org
KEGG108.0 [89]	Metabolic molecular network	http://www.genome.jp/kegg/
WikiPathways [90]	Metabolic molecular network	http://www.wikipathways.org/
Reactome [91]	Biological pathways	http://www.pathwaycommons.org/
PID [92]	Biological pathways	http://pid.nci.nih.gov
Pathway Common [93]	Biological pathways	http://pid.nci.nih.gov
Go 2023-07-27 [94]	Cellular component, molecular function and biological process	http://geneontology.org/

HPRD visually depicts and integrates information related to the domain structure of each protein in the proteome, posttranslational modifications, interaction networks and disease associations [76]. BioGRID contains 2,045,743 protein and genetic interactions from 76,687 publications, 29,093 chemical interactions and 100 posttranslational modifications of 8257 biological species [77]. STRING serves as a functional enrichment analysis platform for PPI networks, containing data from 5090 organisms, 24.6 million proteins and 2 billion protein interactions [95]. iRefWeb provides information on disease-related proteins, genes and their interactions [79]. MINT covers 607 species and 117,001 protein interactions [81]. IntAct contains approximately 275,000 molecular interactions from over 5000 publications [81]. PINA uses a cluster of interaction modules identified from the protein interaction networks, including terms of gene ontology, the KEGG pathway, the Pfam domain and chemical and genetic perturbations of MSigDB [82]. PhosphoSitePlus contains information about more than 300,000 protein posttranslational modification sites and more than 25,000 protein posttranslational modifications affected by variants [83]. Phospho.ELM is a database of experimentally validated phosphorylation sites in eukaryotic proteins, containing 1703 phosphorylation sites for 556 phosphorylated proteins [84]. PTMcode contains 316,546 modification sites from 69 different posttranslational protein modification types involving more than 100,000 of 19 different eukaryotic proteins, totaling 1.6 million sites and 17 million functional associations [85]. Interactome3D is a Web service for the structural annotation of PPI networks; it can predict a set of proteins or interactions in the organisms based on outcome information [86]. 3did is used as a template for interactions between the two globular domains and for novel domain-peptide interactions [87]. Instruct is a database of high-quality protein interactome networks with 3D structural resolution. It contains data for 6585 individuals, including 644 from *Arabidopsis*, 166 from *Drosophila melanogaster*, 119 from *Mus*, 1273 from *Saccharomyces cerevisiae* and 37 interactions [88].

KEGG is a database that helps in understanding the advanced functional and practical experimental technologies related to biological systems (cells, organisms and ecosystems) using molecular-level information (especially through large-scale molecular datasets generated by genome sequencing and other high-throughput sources) [89]. WikiPathways covers an integrated database of major genes, proteins and small-molecule systems, and it also includes canonical signaling pathways that can represent receptor-binding events, protein complexes, phosphorylation reactions, translocation and transcriptional regulation [90]. Reactome is an open database focusing on signaling, metabolic molecules and their involvement in biological pathways and process relationships. It encompasses various components, including nucleic acids, proteins, complexes, vaccines, anticancer therapeutics and small molecules, forming intricate biological interaction networks [91]. The Pathway Common database comprises 5772 pathways and 2,424,055 interactions [93]. GO covers the current scientific knowledge about gene functions in many different organisms from humans to bacteria. It provides insights into the functions of gene-produced proteins and noncoding RNA molecules [94].

3.3. Data for CDG identification methods

Numerous CDG recognition methods have been proposed based on the aforementioned public databases and biological networks. The data for typical CDG identification methods are presented in Table 4.

Table 4. Data for typical CDG identification methods.

Example		Data type	Biological network	Database
Gene mutation frequency-based methods	MutSigCV [8]	Somatic mutations	None	TCGA
	MuSiC [9]	Somatic mutations	None	TCGA
	OncodriveCLUST [10]	Somatic mutations	None	COSMIC and TCGA
	driverMAPS [13]	Somatic mutations	None	TCGA
Network-based methods	DawnRank [11]	Somatic mutations and gene expression	Reactome, PI, KEGG	TCGA
	DriverNet [14]	Somatic mutations and gene expression	PPI network	TCGA, METABRIC, TN and TCGA HGS
	TieDIE [15]	Somatic mutations and gene expression	WikiPathways	TCGA
	HotNet2 [16]	Somatic mutations	HINT + HI2012	TCGA
	MaxMIF [18]	Somatic mutations	STRING	TCGA
	VarWalker [19]	Somatic mutations	HPRD	[96, 97]
	DyTidriver [20]	Somatic mutations and gene expression	human FIN	TCGA and GEO
	OncoIMPACT [21]	Somatic mutation	[98]	TCGA and CCLE
Coverage and mutually exclusive feature-based methods	RME [22]	Somatic mutations and gene expression data	[98]	Cancer Genome Atlas Data Portal
	CoMEt [23]	Mutation datasets	None	TCGA
	Multi-Dendrix [24]	Somatic mutation		[99]
	Dendrix [12]	Somatic mutation		TCGA
	TiMEx [26]	Somatic mutation		TCGA
	MEMo [27]	Somatic mutations and gene expression	[98]	Affymetrix U133 and Agilent expression platforms
	nCOP [28]	Somatic mutation		TCGA

Continued on next page

Example		Data type	Biological network	Database
Machine learning methods	20/20+ [30]	Somatic mutation		TCGA [100]
	DriverML [31]	Cancer driver mutations, passenger mutations and gene expression and copy number variation		TCGA and GDC [101–103]
	Agajanian etc. [3]	Cancer driver mutations and passenger mutations		[104,105]
	Moonlight [32]	Gene expression, methylation, copy number, chromatin accessibility, clinical, mutation and cell lines data		https://www.cancerrxgene.org/downloads
	DeepDriver [33]	Somatic mutation and gene expression		GDC and CGC [106]
	GUST [34]	Mutation data		CGC [106]
	DORGE [35]	Somatic mutation and population genetics		TCGA and COSMIC
		Epigenetic datasets		
		DNA methylation information		gnomAD v2
	CancerMine [37]	Somatic mutation and gene expression		COSMIC
		TSGs and OGs		CGC
	LOTUS [38]	Somatic mutations		COSMIC and TCGA (http://cancergenome.nih.gov/)
		Somatic mutations for comparing with DiffMut and 20/20+		[44]
		Somatic mutations for comparing MutSigCV		GenePattern
		TUSON train set for training		[101]
		20/20 train set for training		[30]
		CGCv86 train set for training		COSMIC
			PPI network	HPRD

4. Algorithms used in computational methods for CDG identification

Effective computational methods can enhance the accuracy of CDG identification and contribute to discovering more novel CDGs. The algorithms used in CDG identification methods are a critical

area of focus. Table 5 displays the common algorithms used in these methods and their corresponding CDG identification methods.

Table 5. Algorithms used in computational methods for CDG identification.

Algorithm	Computational method	Description
PageRank	DawnRank [46]	The PageRank algorithm was used to rank the genes in the gene interaction network
Decision tree	20/20+ [63]	Other ratio measures used mutational function impact bias, mutation clustering pattern, or mutation composition pattern, unlike the ratio 20/20 rule assessing the proportion of inactivating and repeated missense mutations in the gene of interest
SVM		Ninety-five features were obtained from 10 functional impact-based algorithms, and SVM models were trained to predict missense mutations
Random forest	CanDrA [90]	Eighty-six features were used to identify missense mutations with tumor cell proliferation functions
SVM	CHASM [91]	Integrated Rao's score testing and supervised machine learning to identify CDGs
Random forest classifiers and deep convolutional neural networks	DriverML [64]	Integrated different machine learning methods, including tree-based methods, random forest and gradient enhanced tree (GBT) classifiers, and networks with deep convolutional nerves used to predict cancer driver mutations in genomic datasets
Statistical models, hidden Markov models	Agajanian etc. [65]	The method explicitly simulated positive selection at the single-base level and the highly heterogeneous background mutation process. In particular, the selection model used multiple spatial clustering of external annotations and mutations to capture high mutation rates at functionally important loci
Convolutional neural network	driverMAPS [45]	Convolved the mutational features of the genes and their neighbors in the similarity network
Network control strategy	DeepDriver [67]	It was found that driver mutations could drive regulatory networks from normal to disease state
Entropy-based methods	SCS [92]	Subcellular localization information and variant frequency were mutually exclusive in the network
Statistical method	EntroRank [62]	Identified combinations of changes between individuals exhibiting mutually exclusive patterns in the same path, including an exact statistical test of mutual exclusivity, to analyze multiple sets of mutually exclusive and subtype-specific alterations
Markov chain Monte Carlo	CoMEt [55]	High-weight gene sets were sampled using the Markov chain Monte Carlo algorithm
Statistical machine learning methods	Dendrix [58]	Statistical machine learning methods were used to select subsets of genes, and modular network analysis methods were used to identify potential candidate driver genes

Gene mutation frequency-based methods were among the early CDG identification methods used to analyze mutation frequencies. These methods used statistical algorithms to identify significant differences between CDGs and passenger genes. Some of these methods included HotNet2 [16], DriverNet [14], InVEx, OncodriveCLUST [10], MutSigCV [8] and MuSiC [9]. Although these methods effectively identify CDGs with high mutation frequencies, they may perform poorly for genes with low mutation frequencies. In contrast, dNdScv [107] is a suite of maximum-likelihood dN/dS methods designed to quantify selection in cancer and somatic evolution.

Network-based methods usually use network analysis algorithms, such as random walk [11,19], algorithms of complex network feature calculation, greedy algorithm, gravity model [18], belief propagation algorithm [43] and so on.

Coverage and mutually exclusive feature-based methods mainly employ the mutual exclusivity and coverage of driver genes in signaling pathways, integer linear programming [24,28], greedy algorithms [12,14], Cancer3D [69], probabilistic model [26], correlation analysis and statistical analyses [27].

Machine learning methods are employed in CDG identification [108,109]. These methods use kernel techniques to find the optimal classification surfaces between different sample categories, maximizing the interval between them. CDG identification methods such as CanDrA [104], SVMerge [110] and DriverML [18] predict well using SVM. These methods effectively address the issues related to sample distribution and redundant features in driver genes, especially mitigating overfitting problems. However, these methods can be computationally expensive, limiting their scalability to large datasets. Random forest, which is composed of multi-decision trees generated through a bootstrapped resampling technique, is employed by methods such as CHASM [111] and Agajanian [19]. This approach offers benefits like the quantification of feature importance and rapid processing, especially beneficial for datasets with partial data loss. Deep learning methods delve into the intrinsic rules and representation levels of the CDG sample data for interpreting the CDG data, reducing the need for human analysis. Deep learning is a complex machine learning algorithm that far exceeds previously relevant techniques in many ways. Agajanian [3] and DeepDriver [33] employ deep CNNs to achieve CDG identification. In addition, methods such as FATHMM [112], which uses a Bayesian approach, and that suggested by Lu et al. [113], based on Bayesian algorithms, have been employed for CDG identification. Markov models involve dual stochastic processes with hidden Markov chains representing certain states and a set of displayed random functions. Although these models effectively address the labeling issues, they may introduce labeling biases based on homogeneous Markov and observed dividend hypotheses. Examples include Dendrix [12] and PageRank, the latter being a form of Markov chain used by Google to determine the order of search results, and methods such as DawnRank [11].

The aforementioned algorithms are widely used in CDG identification. In the latest research on computational methods, novel algorithms have been introduced: The MaxMIF [18] method is based on a heavy mechanical model; Dendrix [12] and DriverNet [14] use a greedy algorithm; and Multi-Dendrix [24] and Cancer3D [69] use integer linear planning techniques. Heat has also been explored as a CDG identification method [114].

5. A comparative study across eight types of cancer using nine computational methods

We selected nine representative computational methods published since 2014 to investigate the

performances of the four categories of CDG identification methods. These methods were compared and analyzed using validated CDGs associated with eight different cancer types. The performances were evaluated by cumulative number analysis with published CDGs.

The data used in this study comprised validated CDGs and a feature dataset, including mutation data, biological network information and gene expression levels. The validated CDG datasets were extracted from the CGC of Gao et al. (https://cancer.sanger.ac.uk/census/#cl_search). They contained CDGs from eight different types of cancer: kidney chromophobe (KICH), skin cutaneous melanoma (SKCM), breast invasive carcinoma (BRCA), acute myeloid leukemia (LAML), thyroid carcinoma (THCA), glioblastoma multiforme, lung squamous cell carcinoma (LUAD) and uterine corpus endometrial carcinoma (UCEC). Mutation data and gene expression profiles of eight types of cancer were downloaded from TCGA, and STRING was used as the biological network [78].

This study compared the performances of nine computational methods from four different types of CDG identification methods across eight different types of cancer: (1) frequency-based methods (driverMAPS), (2) network-based algorithms (HotNet2, MaxMIF, DNsum, DNmax and OncoIMPACT), (3) coverage and mutually exclusive feature-based method (nCOP) and (4) machine learning-based methods (MutPanning and DriverML). The computational methods that identified more than three CDGs were considered.

Each computational method was used to predict eight cancers. The top 50 genes from these predictions were then analyzed to determine the number of CDG duplications for all 8 cancers using the CGC database. The cumulative numbers of known CDGs from the CGC dataset that were recovered from among the top 50 candidate genes for the 8 cancer types are presented in Table 6 and Figure 1.

Table 6. Cumulative numbers of known CDGs from the CGC dataset recovered from among the top 50 candidate genes across all 8 cancer types.

Method	Number	Method	Number	Method	Number
DNmax	44	DriverML	57	MutPanning	66
DNsum	46	HotNet2	28	OncoIMPACT	63
driverMAPS	5	MaxMIF	66	nCOP	37

The test data underwent prediction and sorting through nine computational methods. The analysis focused on the top 50 candidate genes, selecting the top 5 computational methods with the highest number of identifications. The cumulative numbers of known CDGs from the CGC dataset were recovered from the top 50 candidate genes (Figure 1).

Venn diagrams of 6 computational methods identified 125 CDGs. Among these, 13 CDGs were shared by 5 methods, indicating that 10.4% of CDGs can be identified by these 5 computational methods across all 8 types of cancer. Additionally, 16 CDGs were shared by 4 methods, 20 CDGs were shared by 3 methods, 32 CDGs were shared by 2 methods, and 45 CDGs were shared by 1 method.

The cumulative results suggested that the nine computational methods can be ranked in the following order based on their identification performance: MaxMIF, MutPanning, OncoIMPACT, DriverML, DNsum, DNmax, nCOP, HotNet2 and driverMAPS. Notably, MaxMIF, OncoIMPACT, DNsum and DNmax are network-based algorithms, whereas MutPanning and DriverML belong to the machine learning-based methods. It is evident that both the network-based algorithms and machine learning-based methods exhibited superior identification performance in this analysis.

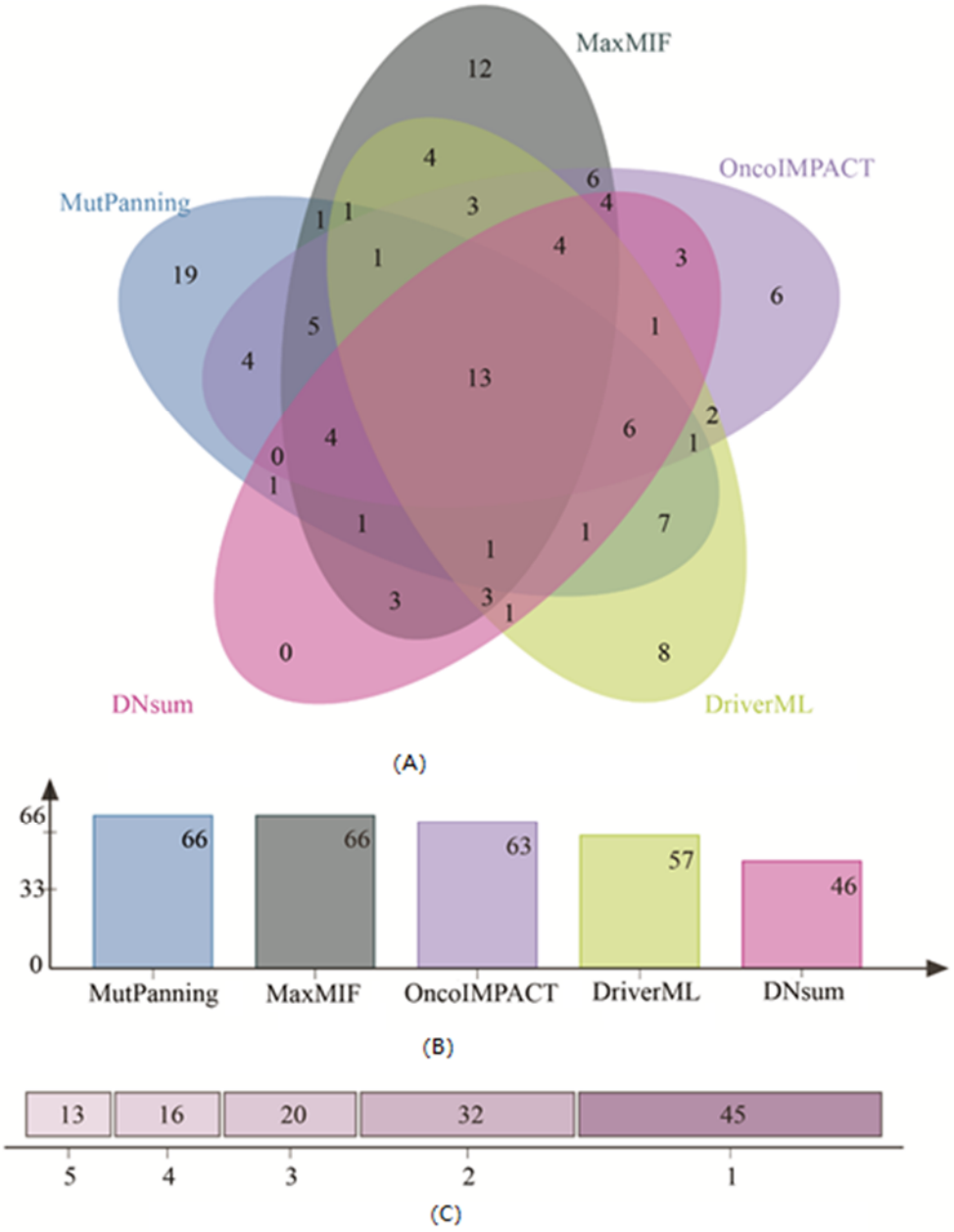

Figure 1. Cumulative numbers of known CDGs from the CGC dataset recovered from the top 50 candidate genes across 8 cancer types. (A) Overlapped CDG numbers were predicted using five CDG identification methods across all eight cancer types. (B) Top 50 CDGs identified by each of the 5 computational methods were selected to calculate the cumulative number of overlaps with the known CDGs in the CGC dataset across 8 cancer types. (C) Number of overlapped genes predicted by n number of CDG identification methods (n = 1, 2, 3, 4, 5). For example, 13 overlapping CDGs were identified using 5 computational methods simultaneously, and 16 overlapping CDGs were identified using 4 computational methods.

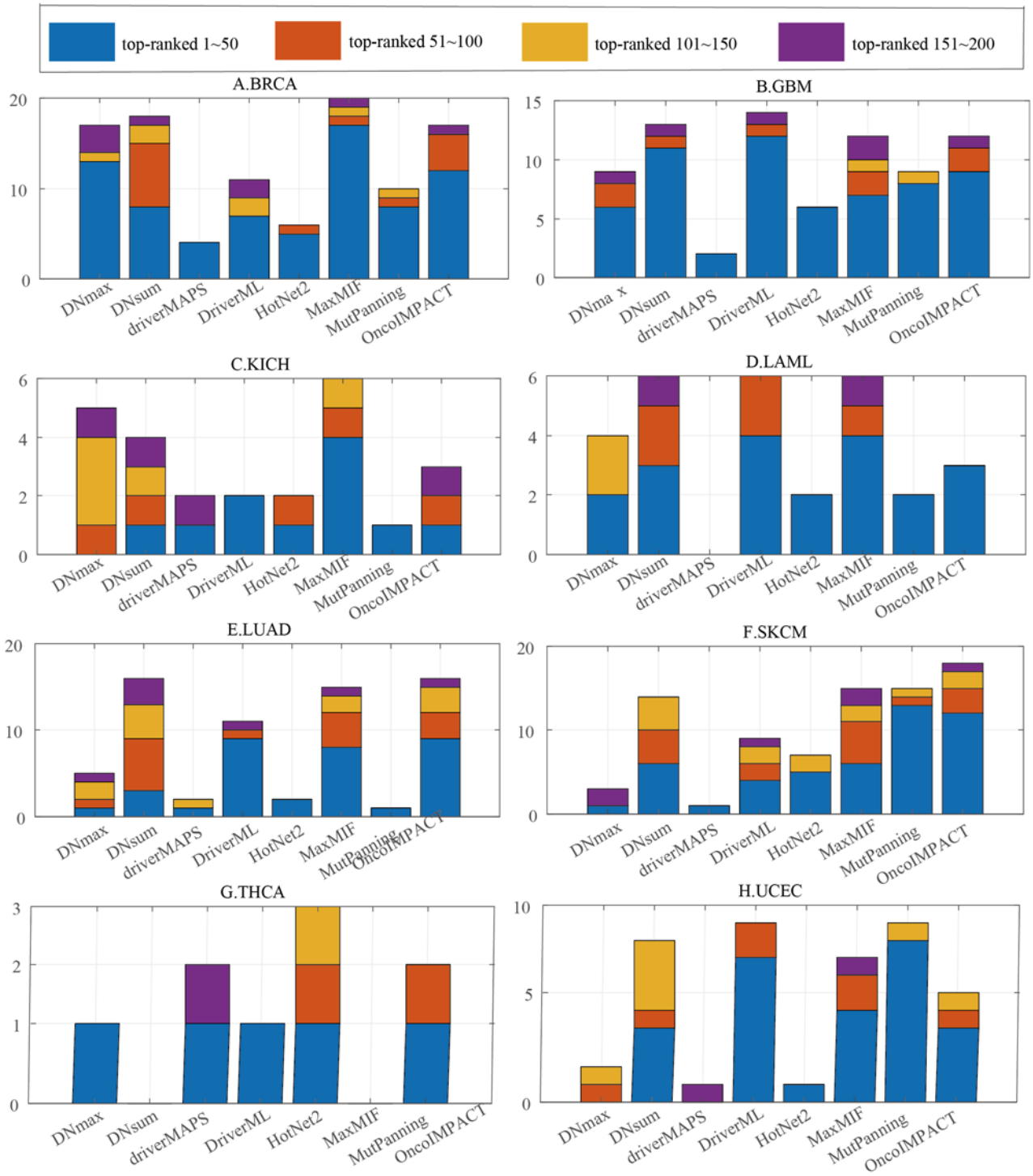

Figure 2. Cumulative numbers of known CDGs from the CGC dataset recovered from top-ranked candidate genes across all eight cancer types.

Figure 2 shows eight algorithms specific for CDG identification in different cancer types. When MaxMIF was used to identify CDGs in BRCA, the cumulative numbers of known CDGs from the CGC dataset recovered in the top 50, 100, 150 and 200 candidate genes were 17, 18, 19 and 20, respectively. Different algorithms exhibited specific strengths in terms of CDG identification. MaxMIF performed best in CDG identification for BRCA, KICH and THCA, while OncoIMPACT performed

well in CDG identification for LUAD and SKCM. Additionally, DriverML exhibited the best performance in terms of CDG identification for GBM, LAML and UCEC. These findings suggested that network-based algorithms and machine learning-based methods consistently outperformed other approaches in this study.

6. Challenges and outlook

This study extensively reviewed biological knowledge, focusing on typical algorithms and software related to CDG research. It also covered various computational methods used to identify CDG-related data resources such as biological networks. Despite significant research efforts establishing a robust foundation for CDG identification, several bottlenecks and technical challenges persist.

(1) Although some CDGs exhibit high-frequency mutations (>20%), most cancer mutations occur at intermediate frequencies (2–20%) or even lower frequencies [8]. For instance, an analysis of 183 lung adenocarcinoma samples revealed that 15% of patients lacked mutations in known cancer genes, highlighting the heterogeneity of cancer mutations [96]. Because of the heterogeneity of cancer mutations, high-frequency mutant CDGs are more vividly identified, while low-frequency genes hold significant potential for exploration. Some studies focus on individual samples or specific cancer types, inadvertently neglecting low-frequency CDGs due to limited sample coverage.

(2) Using the biological network modeling method enhances the possibility of identifying low-frequency CDGs. While the adoption of coverage and mutual exclusivity features is beneficial, the approach to identifying network modules often affects the nodal degree of the network. This can result in prioritizing nodes with a high number of connections [115].

(3) The significance of edges in biological networks is often overlooked. However, cancer omics data provide sufficient resources for common modeling based on both the nodes and edges of these networks. The correlation between adjacent nodes of CDGs represents their perturbative role in gene network pathways. This aspect is not adequately explored in many existing methods.

In conclusion, research on driver gene identification holds significant importance in understanding the function of driver genes. A comprehensive and multi-faceted approach can be undertaken through computational methods, paving the way for advancements before progress in experimental techniques. It is envisaged that computational methods for identifying CDGs can develop in the following directions:

(1) Using a multiomics data integration method for CDG identification represents a crucial step forward. Tumor occurrence and the developmental process rely not solely on a single system. The development of multiomics has improved our access to a large amount of omics information, encompassing protein networks, gene function annotation databases, gene expression profiles and miRNA expression profiles. The multiomics data integration method demonstrates strong consistency and performance, albeit requiring a large sample size to achieve relatively high sensitivity. By considering the expression of neighboring genes, this method effectively identifies CDGs with low frequency. Moreover, the multiomics data integration method is instrumental in capturing specific cancer characteristic signals, facilitating the discovery of specific CDGs at the multiomics level. The surge in multiomics data availability serves as a significant driving force for the advancement of this approach.

(2) Research in computational methods for identifying CDGs should focus on specific types and individualized approaches. However, the specificity of cancer indicates substantial variations in gene characteristics across different types. For example, genes such as EGFR, ALK and MET serve as driver

genes in lung cancer, but in other cancer types, they may not hold the same status. Therefore, developing computational methods tailored to the identification of CDGs specific to certain cancers proves more effective.

(3) The development of algorithms for quantifying the features of tumor heterogeneity is a crucial research area in identifying low-frequency CDGs. Tumor heterogeneity poses a significant challenge in CDG identification, primarily manifesting in two aspects. First, the number of mutations varies considerably among different samples of the same cancer type. Second, the number of gene mutations differs significantly between various tumor samples, with variations up to 100 times between different cancer types. In heterogeneous tumors, a large number of samples exhibit only a few CDG mutations, while a small number of samples contain a large number of CDG mutations. The genome-wide survey showed that the "long tail" phenomenon occurred in the genome frequency distribution, with most CDGs exhibiting low population frequencies. Therefore, devising calculation methods based on mutation samples to quantify mutation characteristics represents one of the current challenges in this field.

As specific computer algorithms and software may not be accessible, showcasing comparative results using a broader range of algorithms is essential. In addition, leveraging biometrics of cancer drivers identified through previous studies or biological experiments is crucial for enhancing the identification performance.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

The project was partially supported by the Humanities and Social Science Fund of the Ministry of Education of China (21YJAZH091), the National Natural Science Foundation (NNSF) of China (6237010008), the Philosophy and Social Science Research Project in Jiangsu Universities (2021SJA1426), The 14th Five-Year Plan of Education Science in Jiangsu Province (T-c/2021/93), the "Textile Light" Higher Education Teaching Reform Research Project of the China National Textile Industry Federation (2021BKJGLX206), the Changzhou science and technology project (CZ20230028, CJ20220151) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (23KJB520002).

Conflicts of interest

The authors declare there is no conflict of interest.

References

1. B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, K. W. Kinzler, Cancer genome landscapes, *Science*, **339** (2013), 1546–1558. https://doi.org/10.1126/science.1235122
2. J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, et al., Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, *Sci. Signal*, **6** (2013), l1. https://doi.org/10.1126/scisignal.2004088

3. S. Agajanian, O. Oluyemi, G. M. Verkhivker, Integration of random forest classifiers and deep convolutional neural networks for classification and biomolecular modeling of cancer driver mutations, *Front. Mol. Biosci.*, **6** (2019), 44. https://doi.org/10.3389/fmolb.2019.00044
4. M. I. Klein, V. L. Cannataro, J. P. Townsend, D. F. Stern, H. Zhao, Identifying combinations of cancer drivers in individual patients, *bioRxiv*, (2019), 674234. https://doi.org/10.1101/674234
5. F. Cheng, J. Zhao, Z. Zhao, Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes, *Briefings Bioinf.*, **17** (2016), 642–656. https://doi.org/10.1093/bib/bbv068
6. W. F. Guo, S. W. Zhang, T. Zeng, T. Akutsu, L. Chen, Network control principles for identifying personalized driver genes in cancer, *Briefings Bioinf.*, **21** (2020), 1641–1662. https://doi.org/10.1093/bib/bbz089
7. M. Sinkala, Mutational landscape of cancer-driver genes across human cancers, *Sci. Rep.*, **13** (2023), 12742. https://doi.org/ARTN 1274210.1038/s41598-023-39608-2
8. M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, et al., Discovery and saturation analysis of cancer genes across 21 tumour types, *Nature*, **505** (2014), 495–501. https://doi.org/10.1038/nature12912
9. N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, et al., MuSiC: identifying mutational significance in cancer genomes, *Genome Res.*, **22** (2012), 1589–1598. https://doi.org/10.1101/gr.134635.111
10. D. Tamborero, A. Gonzalez-Perez, N. Lopez-Bigas, OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes, *Bioinformatics*, **29** (2013), 2238–2244. https://doi.org/10.1093/bioinformatics/btt395
11. J. P. Hou, J. Ma, DawnRank: discovering personalized driver genes in cancer, *Genome Med.*, **6** (2014), 56. https://doi.org/10.1186/s13073-014-0056-8
12. F. Vandin, E. Upfal, B. J. Raphael, De novo discovery of mutated driver pathways in cancer, *Genome Res.*, **22** (2012), 375–385. https://doi.org/10.1101/gr.120477.111
13. S. Zhao, J. Liu, P. Nanga, Y. Liu, A. E. Cicek, N. Knoblauch, et al., Detailed modeling of positive selection improves detection of cancer driver genes, *Nat. Commun.*, **10** (2019), 3399. https://doi.org/10.1038/s41467-019-11284-9
14. A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, et al., DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer, *Genome Biol.*, **13** (2012), R124. https://doi.org/10.1186/gb-2012-13-12-r124
15. E. O. Paull, D. E. Carlin, M. Niepel, P. K. Sorger, D. Haussler, J. M. Stuart, Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE), *Bioinformatics*, **29** (2013), 2757–2764. https://doi.org/10.1093/bioinformatics/btt471
16. M. D. Leiserson, F. Vandin, H. T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, et al., Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes, *Nat. Genet.*, **47** (2015), 106–114. https://doi.org/10.1038/ng.3168
17. A. Cho, J. E. Shim, E. Kim, F. Supek, B. Lehner, I. Lee, MUFFINN: cancer gene discovery via network analysis of somatic mutation data, *Genome Biol.*, **17** (2016), 129. https://doi.org/10.1186/s13059-016-0989-x
18. Y. Hou, B. Gao, G. Li, Z. Su, MaxMIF: A new method for identifying cancer driver genes through effective data integration, *Adv. Sci.*, **5** (2018), 1800640. https://doi.org/10.1002/advs.201800640

19. P. Jia, Z. Zhao, VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data, *PLoS Comput. Biol.*, **10** (2014), e1003460. https://doi.org/10.1371/journal.pcbi.1003460
20. J. Song, W. Peng, F. Wang, J. Wang, Identifying driver genes involving gene dysregulated expression, tissue-specific expression and gene-gene network, *BMC Med. Genomics*, **12** (2019), 168. https://doi.org/10.1186/s12920-019-0619-z
21. D. Bertrand, K. R. Chng, F. G. Sherbaf, A. Kiesel, B. K. Chia, Y. Y. Sia, et al., Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles, *Nucleic Acids Res.*, **43** (2015), e44. https://doi.org/10.1093/nar/gku1393
22. C. A. Miller, S. H. Settle, E. P. Sulman, K. D. Aldape, A. Milosavljevic, Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors, *BMC Med. Genomics*, **4** (2011), 34. https://doi.org/10.1186/1755-8794-4-34
23. M. D. Leiserson, H. T. Wu, F. Vandin, B. J. Raphael, CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer, *Genome Biol.*, **16** (2015), 160. https://doi.org/10.1186/s13059-015-0700-7
24. M. D. Leiserson, D. Blokh, R. Sharan, B. J. Raphael, Simultaneous identification of multiple driver pathways in cancer, *PLoS Comput. Biol.*, **9** (2013), e1003054. https://doi.org/10.1371/journal.pcbi.1003054
25. S. Cristea, J. Kuipers, N. Beerenwinkel, pathTiMEx: Joint inference of mutually exclusive cancer pathways and their progression dynamics, *J. Comput. Biol.*, **24** (2017), 603–615. https://doi.org/10.1089/cmb.2016.0171
26. S. Constantinescu, E. Szczurek, P. Mohammadi, J. Rahnenfuhrer, N. Beerenwinkel, TiMEx: a waiting time model for mutually exclusive cancer alterations, *Bioinformatics*, **32** (2016), 968–975. https://doi.org/10.1093/bioinformatics/btv400
27. G. Ciriello, E. Cerami, C. Sander, N. Schultz, Mutual exclusivity analysis identifies oncogenic network modules, *Genome Res.*, **22** (2012), 398–406. https://doi.org/10.1101/gr.125567.111
28. B. H. Hristov, M. Singh, Network-based coverage of mutational profiles reveals cancer genes, *Cell Syst.*, **5** (2017), 221–229. https://doi.org/10.1016/j.cels.2017.09.003
29. J. Song, W. Peng, F. Wang, An entropy-based method for identifying mutual exclusive driver genes in cancer, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **17** (2020), 758–768. https://doi.org/10.1109/TCBB.2019.2897931
30. C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, R. Karchin, Evaluating the evaluation of cancer driver genes, *Proc. Natl. Acad. Sci. U.S.A.*, **113** (2016), 14330–14335. https://doi.org/10.1073/pnas.1616440113
31. Y. Han, J. Yang, X. Qian, W. C. Cheng, S. H. Liu, X. Hua, et al., DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies, *Nucleic Acids Res.*, **47** (2019), e45. https://doi.org/10.1093/nar/gkz096
32. A. Colaprico, C. Olsen, M. H. Bailey, G. J. Odom, T. Terkelsen, T. C. Silva, et al., Interpreting pathways to discover cancer driver genes with moonlight, *Nat. Commun.*, **11** (2020), 69. https://doi.org/10.1038/s41467-019-13803-0
33. P. Luo, Y. Ding, X. Lei, F. X. Wu, deepDriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks, *Front. Genet.*, **10** (2019), 13. https://doi.org/10.3389/fgene.2019.00013

34. P. Chandrashekar, N. Ahmadinejad, J. Wang, A. Sekulic, J. B. Egan, Y. W. Asmann, et al., Somatic selection distinguishes oncogenes and tumor suppressor genes, *Bioinformatics*, **36** (2020), 1712–1717. https://doi.org/10.1093/bioinformatics/btz851
35. J. Lyu, J. J. Li, J. Su, F. Peng, Y. E. Chen, X. Ge, et al., DORGE: Discovery of oncogenes and tumor suppressor genes using genetic and epigenetic features, *Sci. Adv.*, **6** (2020). https://doi.org/10.1126/sciadv.aba6784
36. M. Sudhakar, R. Rengaswamy, K. Raman, Novel ratio-metric features enable the identification of new driver genes across cancer types, *Sci. Rep.*, **12** (2022), 5. https://doi.org/10.1038/s41598-021-04015-y
37. J. Lever, E. Y. Zhao, J. Grewal, M. R. Jones, S. J. M. Jones, CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer, *Nat. Methods*, **16** (2019), 505–507. https://doi.org/10.1038/s41592-019-0422-y
38. O. Collier, V. Stoven, J. P. Vert, *LOTUS:* A single- and multitask machine learning algorithm for the prediction of cancer driver genes, *PLoS Comput. Biol.*, **15** (2019), e1007381. https://doi.org/10.1371/journal.pcbi.1007381
39. J. Reimand, G. D. Bader, Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers, *Mol. Syst. Biol.*, **9** (2013), 637. https://doi.org/10.1038/msb.2012.68
40. L. Qu, Z. Wang, H. Zhang, Z. Wang, C. Liu, W. Qian, et al., The analysis of relevant gene networks based on driver genes in breast cancer, *Diagnostics*, **12** (2022), 2882. https://doi.org/10.3390/diagnostics12112882
41. X. Shi, H. Teng, L. Shi, W. Bi, W. Wei, F. Mao, et al., Comprehensive evaluation of computational methods for predicting cancer driver genes, *Briefings Bioinf.*, **23** (2022), bbab548. https://doi.org/10.1093/bib/bbab548
42. A. C. Gumpinger, K. Lage, H. Horn, K. Borgwardt, Prediction of cancer driver genes through network-based moment propagation of mutation scores, *Bioinformatics*, **36** (2020), i508–i515. https://doi.org/10.1093/bioinformatics/btaa452
43. S. Ng, E. A. Collisson, A. Sokolov, T. Goldstein, A. Gonzalez-Perez, N. Lopez-Bigas, et al., PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis, *Bioinformatics*, **28** (2012), i640–i646. https://doi.org/10.1093/bioinformatics/bts402
44. K. Shi, L. Gao, B. Wang, Discovering potential cancer driver genes by an integrated network-based approach, *Mol. Biosyst.*, **12** (2016), 2921–2931. https://doi.org/10.1039/c6mb00274a
45. C. Suo, O. Hrydziuszko, D. Lee, S. Pramana, D. Saputra, H. Joshi, et al., Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival, *Bioinformatics*, **31** (2015), 2607–2613. https://doi.org/10.1093/bioinformatics/btv164
46. E. Hodzic, R. Shrestha, K. Zhu, K. Cheng, C. C. Collins, S. Cenk Sahinalp, Combinatorial detection of conserved alteration patterns for identifying cancer subnetworks, *Gigascience*, **8** (2019), giz024. https://doi.org/10.1093/gigascience/giz024
47. E. Lusito, B. Felice, G. D'Ario, A. Ogier, F. Montani, P. P. Di Fiore, et al., Unraveling the role of low-frequency mutated genes in breast cancer, *Bioinformatics*, **35** (2018), 36–46. https://doi.org/10.1093/bioinformatics/bty520
48. F. Li, L. Gao, X. Ma, X. Yang, Detection of driver pathways using mutated gene network in cancer, *Mol. Biosyst.*, **12** (2016), 2135–2141. https://doi.org/10.1039/C6MB00084C
49. B. Gao, G. Li, J. Liu, Y. Li, X. Huang, Identification of driver modules in pan-cancer via coordinating coverage and exclusivity, *Oncotarget*, **8** (2017), 36115–36126. https://doi.org/10.18632/oncotarget.16433

50. D. Silverbush, S. Cristea, G. Yanovich-Arad, T. Geiger, N. Beerenwinkel, R. Sharan, Simultaneous integration of multi-omics data improves the identification of cancer driver modules, *Cell Syst.*, **8** (2019), 456–466 e5. https://doi.org/10.1016/j.cels.2019.04.005
51. A. Garavand, C. Salehnasab, A. Behmanesh, N. Aslani, A. H. Zadeh, M. Ghaderzadeh, Efficient model for coronary artery disease diagnosis: a comparative study of several machine learning algorithms, *J. Healthcare Eng.*, **2022** (2022), 5359540. https://doi.org/10.1155/2022/5359540
52. S. J. Malebary, Y. D. Khan, Evaluating machine learning methodologies for identification of cancer driver genes, *Sci. Rep.*, **11** (2021), 12281. https://doi.org/10.1038/s41598-021-91656-8
53. S. W. Zhang, Z. N. Wang, Y. Li, W. F. Guo, Prioritization of cancer driver gene with prize-collecting steiner tree by introducing an edge weighted strategy in the personalized gene interaction network, *BMC Bioinf.*, **23** (2022), 341. https://doi.org/10.1186/s12859-022-04802-y
54. P. H. Acosta, V. Panwar, V. Jarmale, A. Christie, J. Jasti, V. Margulis, et al., Intratumoral resolution of driver gene mutation heterogeneity in renal cancer using deep learning, *Cancer Res.*, **82** (2022), 2792–2806. https://doi.org/10.1158/0008-5472.CAN-21-2318
55. F. Sadoughi, M. Ghaderzadeh, A hybrid particle swarm and neural network approach for detection of prostate cancer from benign hyperplasia of prostate, *Stud. Health Technol. Inf.*, **205** (2014), 481–485.
56. A. J. Moshayedi, A. S. Roy, A. Kolahdooz, S. Yang, Deep learning application pros and cons over algorithm, *EAI Endorsed Trans. AI Rob.*, **1** (2022), 1–13
57. M. Gheisari, G. Wang, M. Z. A. Bhuiyan, A survey on deep learning in big data, in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, (2017), 173–180.
58. U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, et al., An integrated approach to uncover drivers of cancer, *Cell*, **143** (2010), 1005–1017. https://doi.org/10.1016/j.cell.2010.11.013
59. Y. Chen, J. Hao, W. Jiang, T. He, X. Zhang, T. Jiang, et al., Identifying potential cancer driver genes by genomic data integration, *Sci. Rep.*, **3** (2013), 3538. https://doi.org/10.1038/srep03538
60. K. M. Jagodnik, Y. Shvili, A. Bartal, HetIG-PreDiG: A heterogeneous integrated graph model for predicting human disease genes based on gene expression, *PLoS One*, **18** (2023), e0280839. https://doi.org/10.1371/journal.pone.0280839
61. Y. Chen, X. Wu, R. Jiang, Integrating human omics data to prioritize candidate genes, *BMC Med. Genomics*, **6** (2013), 57. https://doi.org/10.1186/1755-8794-6-57
62. Z. Tian, M. Guo, C. Wang, L. Xing, L. Wang, Y. Zhang, Constructing an integrated gene similarity network for the identification of disease genes, *J. Biomed. Semant.*, **8** (2017), 32. https://doi.org/10.1186/s13326-017-0141-1
63. L. Chin, J. N. Andersen, P. A. Futreal, Cancer genomics: from discovery science to personalized medicine, *Nat. Med.*, **17** (2011), 297–303. https://doi.org/10.1038/nm.2323
64. R. Edgar, M. Domrachev, A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res.*, **30** (2002), 207–210. https://doi.org/10.1093/nar/30.1.207
65. J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, et al., The international cancer genome consortium data portal, *Nat. Biotechnol.*, **37** (2019), 367–369. https://doi.org/10.1038/s41587-019-0055-9

66. Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium, Pharmacogenomic agreement between two cancer cell line data sets, *Nature*, **528** (2015), 84–87. https://doi.org/10.1038/nature15736
67. J. Pinero, J. M. Ramirez-Anguita, J. Sauch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, et al., The DisGeNET knowledge platform for disease genomics: 2019 update, *Nucleic Acids Res.*, **48** (2020), D845–D855. https://doi.org/10.1093/nar/gkz1021
68. D. Repana, J. Nulsen, L. Dressler, M. Bortolomeazzi, S. K. Venkata, A. Tourna, et al., The Network of Cancer Genes (NCG), a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens, *Genome Biol.*, **20** (2019), 1. https://doi.org/10.1186/s13059-018-1612-0
69. M. Sedova, M. Iyer, Z. Li, L. Jaroszewski, K. W. Post, T. Hrabe, et al., Cancer3D 2.0: interactive analysis of 3D patterns of cancer mutations in cancer subsets, *Nucleic Acids Res.*, **47** (2019), D895–D899. https://doi.org/10.1093/nar/gky1098
70. R. Mosca, J. Tenorio-Laranga, R. Olivella, V. Alcalde, A. Ceol, M. Soler-Lopez, et al., dSysMap: exploring the edgetic role of disease mutations, *Nat. Methods*, **12** (2015), 167–168. https://doi.org/10.1038/nmeth.3289
71. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome, *Nature*, **489** (2012), 57–74. https://doi.org/10.1038/nature11247
72. E. C. Roadmap, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, et al., Integrative analysis of 111 reference human epigenomes, *Nature*, **518** (2015), 317–330. https://doi.org/10.1038/nature14248
73. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, et al., An atlas of active enhancers across human cell types and tissues, *Nature*, **507** (2014), 455–461. https://doi.org/10.1038/nature12787
74. G. T. Consortium, The Genotype-Tissue Expression (GTEx) project, *Nat Genet.*, **45** (2013), 580–585. https://doi.org/10.1038/ng.2653
75. S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, et al., COSMIC: exploring the world's knowledge of somatic mutations in human cancer, *Nucleic Acids Res.*, **43** (2015), D805–D811. https://doi.org/10.1093/nar/gku1075
76. T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, et al., Human protein reference database—2009 update, *Nucleic Acids Res.*, **37** (2009), D767–D772. https://doi.org/10.1093/nar/gkn892
77. A. Chatr-Aryamontri, B. J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, et al., The BioGRID interaction database: 2013 update, *Nucleic Acids Res.*, **41** (2013), D816–D823. https://doi.org/10.1093/nar/gks1158
78. D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, et al., The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets, *Nucleic Acids Res.*, **49** (2021), D605–D612. https://doi.org/10.1093/nar/gkaa1074
79. B. Turner, S. Razick, A. L. Turinsky, J. Vlasblom, E. K. Crowdy, E. Cho, et al., iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence, *Database*, **2010** (2010), baq023. https://doi.org/10.1093/database/baq023

80. L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, et al., MINT, the molecular interaction database: 2012 update, *Nucleic Acids Res.*, **40** (2012), D857–D861. https://doi.org/10.1093/nar/gkr930
81. S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, et al., The IntAct molecular interaction database in 2012, *Nucleic Acids Res.*, **40** (2012), D841–D846. https://doi.org/10.1093/nar/gkr1088
82. M. J. Cowley, M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, et al., PINA v2.0: mining interactome modules, *Nucleic Acids Res.*, **40** (2012), D862–D865. https://doi.org/10.1093/nar/gkr967
83. P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, et al., PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse, *Nucleic Acids Res.*, **40** (2012), D261–D270. https://doi.org/10.1093/nar/gkr1122
84. F. Diella, S. Cameron, C. Gemund, R. Linding, A. Via, B. Kuster, et al., Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins, *BMC Bioinf.*, **5** (2004), 79. https://doi.org/10.1186/1471-2105-5-79
85. P. Minguez, I. Letunic, L. Parca, L. Garcia-Alonso, J. Dopazo, J. Huerta-Cepas, et al., PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins, *Nucleic Acids Res.*, **43** (2015), D494–D502. https://doi.org/10.1093/nar/gku1081
86. R. Mosca, A. Ceol, P. Aloy, Interactome3D: adding structural details to protein networks, *Nat. Methods*, **10** (2013), 47–53. https://doi.org/10.1038/nmeth.2289
87. R. Mosca, A. Ceol, A. Stein, R. Olivella, P. Aloy, 3did: a catalog of domain-based interactions of known three-dimensional structure, *Nucleic Acids Res.*, **42** (2014), D374–D379. https://doi.org/10.1093/nar/gkt887
88. M. J. Meyer, J. Das, X. Wang, H. Yu, INstruct: a database of high-quality 3D structurally resolved protein interactome networks, *Bioinformatics*, **29** (2013), 1577–1579. https://doi.org/10.1093/bioinformatics/btt181
89. M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, Data, information, knowledge and principle: back to metabolism in KEGG, *Nucleic Acids Res.*, **42** (2014), D199–D205. https://doi.org/10.1093/nar/gkt1076
90. T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, et al., WikiPathways: building research communities on biological pathways, *Nucleic Acids Res.*, **40** (2012), D1301–D1307. https://doi.org/10.1093/nar/gkr1074
91. D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, et al., The Reactome pathway knowledgebase, *Nucleic Acids Res*., **42** (2014), D472–D477. https://doi.org/10.1093/nar/gkt1102
92. C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, et al., PID: the pathway interaction database, *Nucleic Acids Res.*, **37** (2009), D674–D679. https://doi.org/10.1093/nar/gkn653
93. E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, et al., Pathway Commons, a web resource for biological pathway data, *Nucleic Acids Res.*, **39** (2011), D685–D690. https://doi.org/10.1093/nar/gkq1039
94. H. Mi, A. Muruganujan, D. Ebert, X. Huang, P. D. Thomas, PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools, *Nucleic Acids Res.*, **47** (2019), D419–D426. https://doi.org/10.1093/nar/gky1038

95. A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, et al., STRING v9.1: protein-protein interaction networks, with increased coverage and integration, *Nucleic Acids Res.*, **41** (2013), D808–D815. https://doi.org/10.1093/nar/gks1094
96. M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, et al., Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing, *Cell*, **150** (2012), 1107–1120. https://doi.org/10.1016/j.cell.2012.08.029
97. E. Hodis, I. R. Watson, G. V. Kryukov, S. T. Arold, M. Imielinski, J. P. Theurillat, et al., A landscape of driver mutations in melanoma, *Cell*, **150** (2012), 251–263. https://doi.org/10.1016/j.cell.2012.06.024
98. G. Wu, X. Feng, L. Stein, A human functional protein interaction network and its application to cancer data analysis, *Genome Biol.*, **11**(2010), R53. https://doi.org/10.1186/gb-2010-11-5-r53
99. *The Cancer Genome Atlas Network*, Comprehensive molecular portraits of human breast tumours, *Nature*, **490** (2012), 61–70. https://doi.org/10.1038/nature11412
100. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, et al., Signatures of mutational processes in human cancer, *Nature*, **500** (2013), 415–421. https://doi.org/10.1038/nature12477
101. T. Davoli, A. W. Xu, K. E. Mengwasser, L. M. Sack, J. C. Yoon, P. J. Park, et al., Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome, *Cell*, **155** (2013), 948–962. https://doi.org/10.1016/j.cell.2013.10.011
102. H. Rizvi, F. Sanchez-Vega, K. La, W. Chatila, P. Jonsson, D. Halpenny, et al., Molecular determinants of response to anti-programmed cell death (PD)-1 and anti-programmed death-ligand 1 (PD-L1) blockade in patients with non-small-cell lung cancer profiled with targeted next-generation sequencing, *J. Clin. Oncol.*, **36** (2018), 633–641. https://doi.org/10.1200/jco.2017.75.3384
103. R. D. Kumar, A. C. Searleman, S. J. Swamidass, O. L. Griffith, R. Bose, Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data, *Bioinformatics*, **31** (2015), 3561–3568. https://doi.org/10.1093/bioinformatics/btv430
104. Y. Mao, H. Chen, H. Liang, F. Meric-Bernstam, G. B. Mills, K. Chen, CanDrA: cancer-specific driver missense mutation annotation with optimized features, *PLoS One*, **8** (2013), e77945. https://doi.org/10.1371/journal.pone.0077945
105. L. G. Martelotto, C. K. Ng, M. R. De Filippo, Y. Zhang, S. Piscuoglio, R. S. Lim, et al., Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations, *Genome Biol.*, **15** (2014), 484. https://doi.org/10.1186/s13059-014-0484-1
106. M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, et al., Comprehensive characterization of cancer driver genes and mutations, *Cell*, **173** (2018), 371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060
107. I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, et al., Universal patterns of selection in cancer and somatic tissues, *Cell*, **173** (2018), 1823. https://doi.org/10.1016/j.cell.2018.06.001
108. R. Andrades, M. Recamonde-Mendoza, Machine learning methods for prediction of cancer driver genes: a survey paper, *Briefings Bioinf.*, **23** (2022). https://doi.org/10.1093/bib/bbac062
109. S. Parvandeh, L. A. Donehower, K. Panagiotis, T. K. Hsu, J. K. Asmussen, K. Lee, et al., EPIMUTESTR: a nearest neighbor machine learning approach to predict cancer driver genes from the evolutionary action of coding variants, *Nucleic Acids Res.*, **50** (2022), e70. https://doi.org/10.1093/nar/gkac215

110. K. Wong, T. M. Keane, J. Stalker, D. J. Adams, Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly, *Genome Biol.*, **11** (2010), R128. https://doi.org/10.1186/gb-2010-11-12-r128
111. H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, et al., Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations, *Cancer Res.*, **69** (2009), 6660–6667. https://doi.org/10.1158/0008-5472.CAN-09-1133
112. H. A. Shihab, J. Gough, D. N. Cooper, I. N. Day, T. R. Gaunt, Predicting the functional consequences of cancer-associated amino acid substitutions, *Bioinformatics*, **29** (2013), 1504–1510. https://doi.org/10.1093/bioinformatics/btt182
113. X. Lu, X. Li, P. Liu, X. Qian, Q. Miao, S. Peng, The integrative method based on the module-network for identifying driver genes in cancer subtypes, *Molecules*, **23** (2018), 183. https://doi.org/10.3390/molecules23020183
114. F. Yuan, X. Cao, Y. H. Zhang, L. Chen, T. Huang, Z. Li, et al., Identification of novel lung cancer driver genes connecting different omics levels with a heat diffusion algorithm, *Front. Cell Dev. Biol.*, **10** (2022), 825272. https://doi.org/10.3389/fcell.2022.825272
115. M. Tsuchiya, M. Tomita, M. Hashimoto, Robust global regulations of gene expression in biological processes: a major driver of cell fate decision revealed, in *2012 ICME International Conference on Complex Medical Engineering (CME)*, (2012), 744–749. https://doi.org/10.1109/ICCME.2012.6275649