Discover Oncology

REVIEW Open Access

Comprehensive review of publicly accessible databases for gastrointestinal cancer and future challenges in database construction

Han Zhang[1] and Shicai Liu[2*]

*Correspondence:
Shicai Liu
liushicainj@163.com
[1]School of Basic Medical Sciences, Wannan Medical College, Wuhu 241002, China
[2]School of Medical Information, Wannan Medical College, No. 22 Wenchang West Road, Wuhu 241002, China

Abstract

Cancer database is a systematic collection and analysis of various human cancer information at the genomic and molecular levels, which can be used to understand the mechanisms of cancer occurrence and development, and facilitate the diagnosis and treatment of cancer. Gastrointestinal (GI) cancer is one of the leading causes of incidence and mortality worldwide. The progress and development of high-throughput sequencing technology have led to the emergence of a large amount of cancer-related omics data, such as genomics, transcriptomics, and proteomics, laying a data foundation for the diagnosis and treatment of cancer. These omics data are stored in various databases. Public resources promote the progress of scientific research. High-throughput sequencing technology and the era of big data have created a need for collaboration and data sharing in order to effectively utilize this new knowledge. This article provided comprehensive information on 54 publicly accessible databases containing GI cancer-related information, including 44 general databases and 10 specific databases designed for GI cancer. Moreover, we discussed the challenges faced in the construction of GI cancer database and future research directions. We hope that such an introduction can raise awareness and promote the full and effective utilization of these data resources, thereby improving the management of patients with GI cancer.

Keywords Database, Gastrointestinal cancer, High-throughput sequencing, Collaboration, Sharing

1 Introduction

Cancer is the main cause of death and an important obstacle to increasing life expectancy worldwide [1]. In 2022, the top 10 cancer types accounted for 63.4% of new cancer cases and 69.7% of cancer deaths. Among the top ten cancer types in terms of new cases and cancer deaths, gastrointestinal (GI) cancer accounts for three and five, respectively. GI cancer accounted for 18.8% of new cancer cases and 33.3% of cancer deaths. Cancer has become a major burden on the healthcare system, and GI cancer is the most common cause of death among cancer patients [2, 3].

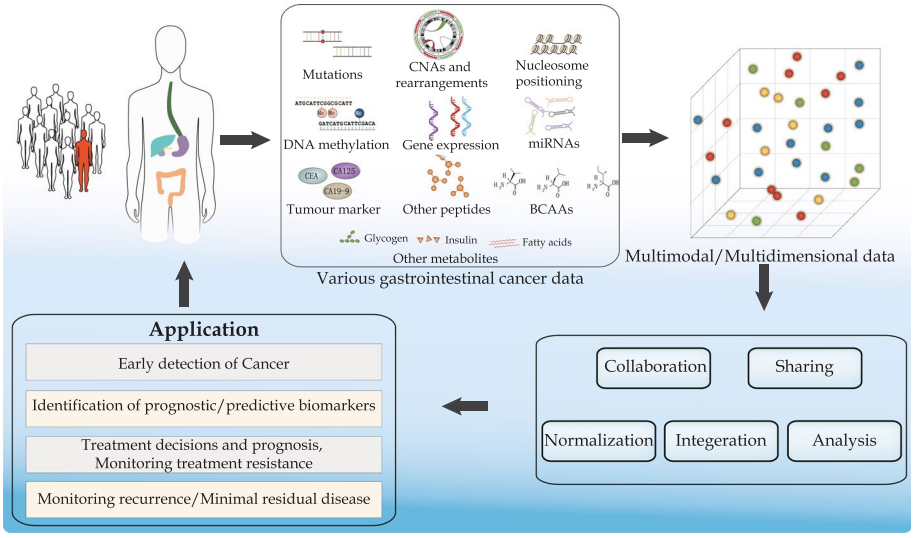

Fig. 1 The data comes from the clinic and serves the clinic

In recent years, advances in high-throughput sequencing technology have enabled the exploration of cancer biology at the genetic and molecular levels, facilitating early detection, personalized diagnosis, and targeted therapy [4–8]. The development of high-throughput sequencing technology has generated massive amounts of omics data, including genomics, transcriptomics, epigenomics, proteomics, and metabolomics data, which provide unprecedented opportunities to decipher the complex mechanisms of GI cancer initiation and progression. Figure 1 clearly presents the workflow and application value of GI cancer data. First, samples were collected from patients with GI cancer, and multi-dimensional omics data (including gene mutations, gene expression, Copy Number Variation (CNV) rearrangements, epigenetic modifications, and metabolites) were integrated from various GI cancer data resources. Following standardization, integration, and analysis, these data form multimodal datasets. Then, through collaboration and sharing mechanisms, data circulation is achieved, ultimately supporting core clinical applications including early cancer detection, prognosis/predictive biomarker identification, treatment decision-making, drug resistance monitoring, and recurrence/minimal residual lesion monitoring. This figure visually demonstrates the full-chain value of multi-omics data from sample collection to clinical translation. However, the full utilization of these data is hindered by challenges such as fragmented data storage, lack of standardization, and limited collaboration and sharing among research teams (Fig. 1).

To address these issues, numerous public databases have been developed to integrate, curate, and disseminate cancer-related data. These databases can be categorized into two main types: general databases (covering multiple cancer types and multi-omics data) and GI cancer-specific databases (focused on one or more GI cancer subtypes) (Supplementary Table 1). In this review, we detailed the core features, data content, and update status of each database, discussed their applications in GI cancer research (e.g., early detection, biomarker identification, treatment resistance monitoring), and analyzed the current challenges and future directions in database construction. This review provides a comprehensive reference for researchers, clinicians, and bioinformaticians working in the field of GI cancer.

1.1 Database selection methodology

To ensure the comprehensiveness and rigor of the database selection, we adopted a systematic approach as follows:

1.1.1 Search strategy

Databases were identified through searches in PubMed and Web of Science using keywords such as "gastrointestinal cancer database", "colorectal cancer omics", "liver cancer biomarker database", "public cancer database", and "precision oncology resource". We also manually searched the "Database Issue" of Nucleic Acids Research and the references from relevant review articles. Inclusion Criteria: (1) Publicly accessible without restricted access (excluding commercial databases); (2) Focus on cancer-related data, with explicit relevance to GI cancers; (3) Provide multi-omics data (genomics, transcriptomics, etc.), clinical data, or functional annotations (e.g., drug sensitivity, biomarker information); (4) Have complete documentation and user guides. Exclusion Criteria: (1) Private or proprietary databases requiring paid subscriptions; (2) Databases with no direct relevance to GI cancers; (3) Duplicate databases (e.g., different versions of the same database were merged, with the latest version prioritized). Quality Assessment: For each included database, we verified (1) Availability of the database (URL accessibility); (2) Update frequency and latest version.

2 General databases

General databases cover multiple cancer types and provide multi-dimensional data resources, including genomics, transcriptomics, epigenomics, proteomics, drug sensitivity data, and clinical data. These databases are widely used in GI cancer research due to their large sample sizes, comprehensive data types, and user-friendly analytical tools.

2.1 Comprehensive cancer databases

Comprehensive cancer databases integrate multi-omics data and clinical information across various cancer types, offering researchers one-stop platforms for data access and analysis. The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) [9], a land mark joint project launched in 2006 by the National Cancer Institute and the National Human Genome Research Institute, has generated over 2.5 petabytes of multi-omics data (genomic, transcriptomic, epigenomic, proteomic) and molecularly characterized more than 20,000 primary cancer and matched normal samples spanning 33 cancer types (including GI cancers). Its unique strength lies in the ability to study cross-omics correlations (e.g., the effects of mutations on expression, methylation-expression relationships) using paired multi-omics data from individual patients; however, TCGA provides relatively raw data that requires basic data analysis capabilities. Nevertheless, its data is widely integrated into user-friendly platforms (e.g., cBioPortal [10], UALCAN [11]) to lower access barriers for researchers without specialized analytical expertise. The official data collection of TCGA was formally completed in 2018, and no new raw samples or sequencing data will be added. However, its data is still undergoing post-processing, standardization, and version updates via the Genomic Data Commons (GDC, https://portal.gdc.cancer.gov/) platform. The latest data version of GDC is Data Release v45.0, released on December 4, 2025.

Database Commons (https://ngdc.cncb.ac.cn/databasecommons/) [12], a global biological database catalog established in 2015, differs fundamentally from other resources in that it does not host raw cancer data but aggregates over 7,000 biological databases from 2,444 institutions across 79 countries/regions (as of December 2025), grouping them into 13 categories and collecting 31 pieces of information per database across four modules (basic, classification, contact, article). It also innovatively uses a z-index combining citation counts and user scores to evaluate database quality and influence, has served over 2.69 million users in 216 countries/regions (as of December 16, 2025), and is endorsed by top journals in the field.

The European Genome-phenome Archive (EGA, https://ega-archive.org/) [13], launched in 2008 as a permanent archive for identifiable genetic, molecular, and phenotypic data, hosts over 4,500 studies including numerous GI cancer-related projects. It boasts diverse data types and strict privacy protection protocols, but its strict access authorization and scrutiny limit data availability, and its large-scale, sensitive datasets demand certain technical capabilities and computational resources for utilization.

The UCSC Cancer Genomics Browser (https://genome.ucsc.edu/) [14] functions as an online analytical platform focused on cancer genomics data visualization and analysis. It aggregates multi-source data into genome location-based tracks to enable multi-level comparisons, supports custom annotation additions, provides genome information for over 80 species, and features a user-friendly interface that facilitates high-throughput dataset exploration, making it particularly suitable for visual comparative analyses of GI cancer genomic data.

The Catalogue of Somatic Mutations in Cancer (COSMIC, https://cancer.sanger.ac.uk/cosmic/) [15], initiated in 2004 as a survey of four cancer-related genes and now updated to v103 (2025) with quarterly updates, is the world's largest curated somatic mutation repository covering the entire human genome and all cancer types. It has been widely applied in GI cancer research to identify mutation hotspots (e.g., BRAF V600E in colorectal cancer, TP53 mutations in gastric cancer) and supports targeted therapy development, with somatic variants linked to stable GRCh37/GRCh38 genome coordinates to enhance interoperability.

The cBio Cancer Genomics Portal (cBioPortal, http://www.cbioportal.org/) [10] serves as a comprehensive platform for accessing, downloading, analyzing, and visualizing cancer genomics data, integrating multi-type genomic data (somatic mutations, copy number alterations, mRNA/miRNA expression, methylation, protein abundance) and datasets from large-scale projects like TCGA and ICGC. Its key advantage lies in enabling multi-dimensional visual analysis (e.g., oncoprints, survival curves) of GI cancer cohorts (colon adenocarcinoma, stomach adenocarcinoma) without requiring advanced technical skills.

The International Cancer Genome Consortium (ICGC, https://icgc.org/) database [16] aims to catalog genomic abnormalities across global cancer types and subtypes, housing 86 projects with approximately 25,000 cancer genomes and providing tools for data visualization, querying, and downloading. Notably, its original Data Portal was retired in June 2024, and researchers should now use its successor, ICGC ARGO (https://icgcargo.org/), for ongoing GI cancer cohort data collection.

The University of Alabama at Birmingham Cancer data analysis Portal (UALCAN, http://ualcan.path.uab.edu/index.html) [11] is an intuitive, interactive resource focused

on cancer omics data analysis. It helps researchers explore gene/target-related information, supports clinical proteomic data analysis (total/phosphorylated proteins) and gene-protein expression analysis in pediatric brain tumors, and serves as a practical tool for deepening the understanding of gene expression patterns in GI cancer.

2.2 Cancer genome databases

Cancer genome databases focus on genomic alterations and their clinical implications. The Clinical Genomic Database (CGD, http://research.nhgri.nih.gov/CGD/) [17], a manually curated online disease database established in 2013, centers on conditions with known genetic causes and prioritizes medically relevant genetic data linked to available interventions (including prevention, surveillance, medical, or surgical treatments). It has seen a 79% increase in the number of clinically organized genes by affected organ system, and currently features 623 manifestation-related genes and 174 intervention-related genes in the gastrointestinal category, with continuous maintenance and updates to ensure data timeliness. By contrast, the Cancer Hotspots database (https://www.cancerhotspots.org/#/home) [18, 19] specializes in statistically significant recurrent mutations identified from large-scale cancer genomics data, currently documenting mutation hotspots for single residues and in-frame indels derived from 24,592 tumor samples. This focus makes it particularly valuable for elucidating cancer onset and progression mechanisms, as well as developing targeted treatment strategies for specific mutations. OncoKB™ (http://oncokb.org) [20, 21], a precision oncology knowledge repository, stands out for providing detailed biological and clinical annotations of cancer genomic alterations; its annotations are accessible via a public web portal and integrated into the cBioPortal for Cancer Genomics, serving as a practical tool for physicians and researchers to interpret genomic changes and guide precision medicine practices.

2.3 Cancer transcriptome databases

Cancer transcriptome databases focus on transcriptomic profiles and regulatory mechanisms. LncTarD 2.0 (http://bio-bigdata.hrbmu.edu.cn/LncTarD) [22], an upgraded version of the comprehensive lncRNA-focused resource, specializes in experimentally supported targeted regulation of key lncRNAs, their functional effects, and regulatory mechanisms in lncRNA-mediated diseases. Compared with its predecessor [23], it expands data coverage and adds practical new features including the prediction of potential diagnostic or therapeutic biomarkers for circulating tumor cells, integration of experimentally validated lncRNA-target interactions, and dedicated tools for RNA-seq/microarray and single-cell analysis of lncRNA-target regulation in diseases, making it a valuable platform for exploring lncRNA functions and mechanisms in cancer pathogenesis and identifying novel biomarkers and therapeutic targets.

In contrast, Gene Expression Profiling Interactive Analysis (GEPIA, http://gepia.cancer-pku.cn/index.html) [24] is a user-friendly online bioinformatics tool that aggregates RNA sequencing data from 9,736 tumor tissues and 8,587 normal tissues sourced from TCGA and GTEx databases. It offers a suite of core functions including gene expression analysis, gene correlation analysis, survival analysis, and dimension reduction analysis, and allows users to upload custom datasets for differential analysis against TCGA/GTEx cancer or normal data. It eliminates the need for programming expertise, thus serving as an accessible resource for cancer gene expression research. Its standalone extension

GEPIA2021 (http://gepia2021.cancer-pku.cn/) [25], launched in 2021, further enables multiple interactive analyses based on deconvolution results to meet more advanced research needs.

GEPIA3 (https://gepia3.bioinfoliu.com/) [26] is the latest upgraded iteration of the GEPIA series, which further expands the platform's analytical capabilities based on the foundation of GEPIA and GEPIA2021, focusing on pan-cancer multi-omics analysis. It adds three innovative functional modules. first is a comprehensive drug sensitivity analysis module that integrates TCGA patient-derived data with cell-line screening data for more than 1,000 therapeutic agents, realizing the association analysis between gene expression and drug response; second is an advanced interaction analysis module, which supports the analysis of gene/protein interactions and regulatory networks across multiple cancer types; third is an RNA alterations analysis module, which incorporates allele-specific expression, alternative promoters and RNA fusions data based on the Pan-Cancer Analysis of Whole Genomes (PCAWG) project. These updates make GEPIA3 a more effective and comprehensive platform, which can help researchers mine large-scale genomic datasets to explore cancer mechanisms, screen therapeutic targets and develop precision medicine strategies.

2.4 Cancer DNA methylation databases

Cancer DNA methylation databases focus on epigenetic modifications of DNA and their relevance to cancer pathogenesis. This study mainly introduces two upgraded platforms that differ in data scope and functional design for targeted research. DiseaseMeth version 3.0 (http://diseasemeth.edbc.org/) [27], an iteration of the human disease methyl ation database following its initial [28] and 2.0 versions [29], expands the coverage of diseases and samples while providing a more comprehensive catalog of disease-related genetic associations. Compared with its earlier versions, it also enhances differential analysis tools and develops new web-based functions for functional annotation, cluster analysis, and survival analysis, making it a practical resource for mining methylation data and exploring the molecular mechanisms of human diseases including GI cancer. By contrast, MethHC 2.0 (https://awi.cuhk.edu.cn/~MethHC/methhc_2020/php/index.php) [30, 31], the updated version of the comprehensive methylation-focused database, specializes in systematically integrating DNA methylation data with mRNA/miRNA expression profiles in human cancers (with a particular focus on GI cancer). It aggregates multi-omics data from multiple public sources, including DNA methylomes and transcriptomes spanning 33 human cancers, over 50,118 microarray and RNA sequencing datasets from TCGA and GEO, as well as 3,586 manually curated entries and experimental evidence extracted from more than 7,000 published studies. It further improves data annotation standards and optimizes its user-friendly web interface to facilitate efficient data query, visualization, and representation, thereby supporting in-depth exploration of the correlations between DNA methylation and gene expression in cancer.

2.5 Cancer proteome databases

Cancer proteome databases focus on protein expression profiles, proteogenomic correlations, and their implications for cancer research. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) [32] is a large-scale collaborative project dedicated to advancing the understanding of cancer's molecular basis through integrated proteomic

and genomic analyses. Coordinated by its Alliance's three specialized centers (Proteome Characterization, Proteome Translational Research, Proteome Data Analysis), CPTAC generates and publicly shares multi-dimensional data (genomic, proteomic, imaging data), along with associated tests and reagents to accelerate cancer research and clinical translation. It has provided genomic data from over 1,500 cancer patients to the GDC platform, with its proteomic datasets accessible via the CPTAC Data Portal and Proteomic Data Commons (PDC).

The Pan-Cancer Proteome Atlas (TPCPA) [33] is a mass spectrometry-based pan-cancer proteome database, built based on data-independent acquisition mass spectrometry technology. Its core value lies in breaking the limitation of single cancer type research in most current cancer proteomics studies. It contains 9,670 proteins derived from 999 primary tumors covering 22 cancer types and provides a dedicated web resource (http://r2platform.com/TPCPA/) for data query. Compared with CPTAC, TPCPA focuses more on the analysis of pan-cancer and cancer-specific protein expression characteristics: it identifies pan-cancer and cancer type-enriched proteins with extensive external annotations, screens for candidate drug targets and biomarkers, such as E3-ubiquitin ligases (HERC5 in esophageal cancer, RNF5 in liver cancer) that can be used for proteolysis-targeting chimeras, and 13 co-expression modules with hub proteins (GFPT1, LRPPRC, PINK1, DOCK2, PTPN6) as potential drug targets; it also identifies protein markers for RNA-based consensus molecular subtypes and two immune subtypes with prognostic value in colorectal cancer, and constructs a cancer type classifier for identifying cancers of unknown primary origin.

2.6 Cancer metabolome databases

Cancer Metabolic Biomarker Database (CMBD) [34] is a curated repository focused on tumor metabolism, compiling cancer metabolic biomarkers identified from PubMed literatures. To date, it includes 438 carefully curated relationships linking 282 biomarkers to 76 cancer subtypes across 18 tissues, including those related to GI cancer. Users can find extensive information on metabolic biomarkers associated with cancer, along with references, clinical samples, and their interconnections within the database. CMBD serves as a valuable resource for advancing the study of cancer's metabolic pathways and is accessible at http://www.sysbio.org.cn/CMBD/.

2.7 Cancer-related gene databases

DriverDBv4 (http://driverdb.bioinfomics.org/) [35], the fourth iteration of the DriverDB series of tumor-related gene databases, builds on the unique value of its earlier versions while expanding its capabilities to support more comprehensive cancer driver gene research, with clear differences in data scale and functional focus across each update. The original DriverDB [36] focused on driver gene/mutation identification, housing 6,079 exceptional exome-sequence datasets, annotation resources, and published bioinformatics algorithms, and providing "cancer" and "gene" dual perspectives to visualize the links between cancer and driver genes/mutations. DriverDBv2 [37] expanded the dataset scale to over 9,500 cancer-associated RNA-seq datasets and over 7,000 exome-seq datasets, adding seven new computational algorithms for driver gene identification, as well as "expression" and "hot spots" functions in the "Genes" section. DriverDBv3 [38] shifted to emphasize concise data visualization for interpreting complex cancer omics

information, integrating computational tools to define CNV and methylation drivers, and adding four new functions (CNV, methylation, survival, and miRNA). DriverDBv4 represents the most comprehensive upgrade, increasing the number of cancer cohorts from 33 to 70 (covering approximately 24,000 samples), integrating proteomic data to expand the scope of omics analysis, adding multiple multi-omics algorithms for cancer driver identification, optimizing visualizations to summarize high-volume data, redesigning sections to accommodate larger datasets, and adding two custom analysis features for multi-omics driver identification and subgroup representation analysis. Compared to its earlier versions, DriverDBv4 is able to support the comprehensive interpretation of multi-omics data across different cancer types, helping to deepen the understanding of cancer heterogeneity and support the development of personalized clinical strategies.

2.8 Cancer and drug databases

Cancer and drug databases focus on the correlation between tumor molecular characteristics and drug response. The Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org) [39], developed by the UK Sanger Institute, is currently the largest public resource for tumor cell drug sensitivity data, and its core value lies in collecting data on the sensitivity and response of tumor cells to drugs. Updated to version 8.5 in October 2023, it added 187 drugs and over 160,000 new unique drug cell line pairs (a 70% increase from the previous version), now containing two main datasets (GDSC1 and the updated GDSC2 from improved Sanger Institute screening), covering the effects of 621 anti-cancer compounds in more than 1,000 cancer cell lines across 24 pathways. Its unique advantage is providing IC50 values for drug sensitivity, which has been applied to develop machine learning models for predicting cetuximab response in GI cancer [40]; while it overlaps with the Cancer Cell Line Encyclopedia (CCLE) in data content, GDSC focuses more specifically on drug sensitivity assays, and its core goal is to identify potential cancer therapeutic targets.

ClinicalOmicsDB (http://trials.linkedomics.org/) [41] differs from GDSC in that it focuses on clinical trial data rather than tumor cell line data. It is built to explore the molecular related to cancer drug responses in clinical settings, housing transcriptomic data from 40 studies covering 11 cancer types (including esophageal cancer NCT03087864 and gastric cancer NCT02589496), with 67 treatment groups covering chemotherapy, targeted therapy, immunotherapy and their combinations. It allows users to explore drug response-related molecular associations through three entry points (studies, treatments, or genes), and connects therapeutic responses with pathway activity and tumor microenvironment characteristics via gene-set analysis, making it more oriented to the translation of preclinical findings to clinical trial scenarios.

3 Specific databases

3.1 Chinese cancer genomic database-esophageal squamous cell carcinoma

The Chinese Cancer Genomic Database-Esophageal Squamous Cell Carcinoma (CCGD-ESCC) [42] is a database that integrates multiple types of genetic association data for esophageal squamous-cell carcinoma (ESCC) in the Chinese population, including genome-wide association studies on ESCC susceptibility in 2,022 ESCC cases and 2,039 normal controls, genome-wide association studies on survival in 1006 ESCC patients,

association studies on genetic variation and gene expression in tumor tissues and paired normal tissues of 94 ESCC patients, and association studies on somatic variation and survival in 675 ESCC patients. This database provides important data resources for genetic and genomic research on ESCC and is freely available at http://esccdb.gao-lab.org/.

3.2 Pancreatic expression database and pancreas genome phenome atlas

Pancreatic Expression Database (PED) [43] and its upgraded iteration Pancreas Genome Phenome Atlas (PGPA) [44] are both dedicated pancreatic disease research resources, with clear differences in positioning and functional focus. PED (http://www.pancreasexpression.org) is a network-based system that has developed into the primary repository for pancreatic-derived omics data, integrating transcriptomic, proteomic, genomic, and miRNA profiles from various pancreas-centric specimens and experimental types. It also functions as a multi-omics data integration and analysis platform, combining a new analytical module with literature mining capabilities, and drawing on data from major public databases (GEO, ArrayExpress) and international projects (TCGA, GENIE, CCLE) to support research tasks including humoral biomarker target discovery, cancer progression-related gene analysis, and cross-platform meta-analysis.

PGPA, the renamed and updated 2025 version of PED, transforms the resource into a centralized, user-friendly portal focused on clinical and molecular data integration for pancreatic diseases. Unlike PED's broader focus on omics data mining, PGPA is powered by SNPNexus (a tool for genetic variation functional annotation) to reduce the analytical burden of large-scale genomic datasets and identify clinically relevant patient variants, and it is closely linked to the UK's national Pancreatic Cancer Research Fund Tissue Bank (PCRFTB) — the only national pancreas tissue bank collecting and storing samples from pancreatic disease patients, their first-degree relatives, and healthy volunteers. PGPA's Analytics Hub has been updated to allow researchers to access and mine both PCRFTB-linked published research data, and large-scale public pancreas-derived datasets, with open, free access (no login required) to support pre-specimen exploration and new discoveries in pancreatic disease research.

3.3 CancerLivER

The Liver Cancer Expression Resource [45] (CancerLivER, https://webs.iiitd.edu.in/raghava/cancerliver/) is a database that compiles datasets on liver cancer gene expression as well as potential biomarkers defined in the literature. Each combined dataset is manually collated and processed using a standardized and uniform pipeline. It contains comprehensive information on 594 biomarkers for liver cancer and is an interactive web platform for searching, browsing and analysis. In addition, the team that developed this database also developed portals such as HCCpred [46] and CancerLSP [47]. HCCpred is a web-bench for predicting non-tumorous and Hepatocellular Carcinoma (HCC) patients based on gene expression profiles. CancerLSP is a web-bench liver cancer stage prediction server based on transcriptomics and epigenomics data.

3.4 Gastric cancer gene database

The Gastric Cancer Gene database [48] (GCGene) is a literature-based genetic resource database designed to assess the priority of genes by their gastric cancer-related

importance and to identify common and unique cellular events at different cancer stages. The database includes 1,815 unique human genes, including 1678 protein-coding genes and 137 non-coding genes, based on an extensive analysis of 3,142 PubMed abstracts. GCGene is available at http://gcgene.bioinfo-minzhao.org/.

3.5 Gastrointestinal cancer database

Gastrointestinal cancer database [49] (GIDB) is a panoramic knowledge database that uses natural language processing methods and multidimensional analysis to automate the management of molecular features. It provides a comprehensive view of approximately 8,700 molecular features mined from over 130,000 texts and 1800 samples, and provides a historical timeline of genes, cancer drug information, as well as molecular-level changes and their impact on prognosis. It features a user-friendly interface and two customizable online tools (heatmaps and networks) for researchers to explore the data effectively. GIDB can be accessed at http://bmtongji.cn/GIDB/index.html.

4 Application of gastrointestinal cancer databases

The application of GI cancer databases is particularly important in modern medical research and clinical practice. These databases provide large-scale data support by integrating clinical, pathological, and genomic data from large numbers of patients, helping researchers conduct in-depth, multi-center studies. This data consolidation not only increases sample sizes and enhances the reliability of statistical analysis, but also facilitates data sharing and drives global collaboration and discovery.

4.1 Early detection of cancer

An important application of GI cancer databases is to facilitate the early detection of cancer. By analyzing risk factors, clinical manifestations, and omics data from different populations, researchers can build early screening models, identify high-risk groups, and develop corresponding screening procedures [50]. Previous studies have shown that based on GI cancer transcriptome data from TCGA, machine learning was used to establish an early cancer diagnosis model and discover new biomarkers [8]. Such early intervention can significantly improve cancer cure rates and reduce mortality caused by late detection. Additionally, these databases can also be used to develop novel screening tools, such as biomarkers based on blood or stool tests [51–53]. For example, Cai et al. [54] constructed a non-invasive early diagnosis model for gastric cancer based on the TCGA database and multi-center clinical samples.

4.2 Identification of prognostic/predictive biomarkers

GI cancer databases can help researchers identify prognostic biomarkers by integrating clinical data, genomic information, and pathological results from large numbers of patients [55–57]. These markers can predict a patient's survival, disease progression, and response to specific treatments. For example, using data from COSMIC, TCGA, GEPIA, cBioPortal, and UALCAN data, researchers identified *TPM4* (tropomyosin 4) gene mutations as predictors of gastric cancer aggressiveness and metastasis [58]. Previous studies have shown that comprehensive analysis of *ING* family genes was conducted using various databases (GEPIA, TCGA, cBioPortal, et al.) to reveal the clinical and biological value of *ING* family genes in liver cancer [59]. Based on multi-omics data

from public databases such as TCGA, Jin et al. [60] developed a gastric cancer prognosis model combined with deep learning, obtaining prognostic features with high predictive value and screening out potential sensitive therapeutic drugs. Tao et al. [61] integrated multi-omics liquid biopsy data and public database resources to systematically explore common molecular markers of GI cancer. Researchers first constructed non-invasive molecular profiles based on 161 clinical plasma samples (including patients with colorectal cancer, gastric adenocarcinoma, and healthy controls) through multi-omics sequencing of cell-free DNA (cfDNA) and cell-free RNA (cfRNA). Subsequently, the association between key molecular features and patient survival outcomes was independently verified using 1,006 GI cancer tissue samples (including colon cancer, rectal cancer, and gastric adenocarcinoma) from the TCGA database. Additionally, by analyzing 38 core cancer-related genes included in the COSMIC database, researchers found that 9 genes, including *CUX1*, *SMAD2*, *QKI*, and *TP53*, showed abnormal changes at the cfDNA or cfRNA level in the plasma of more than 75% of patients with GI cancer. These results suggested that these genes have the potential to serve as common liquid biopsy markers for GI cancer. This strategy effectively integrates real-world liquid biopsy data with authoritative public genomic resources, providing a new path for identifying shared molecular features across cancer types. These markers can not only provide a basis for individualized treatment, but also help doctors assess patients' prognosis at the early stage of treatment, so as to formulate more rational treatment plans.

4.3 Treatment decisions and prognosis, monitoring treatment resistance

By analyzing patients treatment responses and follow-up data from relevant databases, physicians are better able to make treatment decisions and select the most effective treatment options [62]. In addition, these databases can help identify early signs of treatment resistance, such as changes in tumor biomarkers [63], allowing treatment strategies to be adjusted in a timely manner. This dynamic monitoring not only improves the success rate of treatment, but also reduces patients' side effects and enhances the overall therapeutic effect.

4.4 Monitoring recurrence/minimal residual disease

GI cancer databases can provide important support for monitoring cancer recurrence and minimal residual disease (MRD). For example, plasma cfDNA and cfRNA data from the cfOmics database [64] can help to study the biological characteristics of MRD, advance understanding of its formation mechanisms, and provide a theoretical basis for future treatment strategies. Through accurate monitoring, patients can improve their quality of life and prolong their survival time.

4.5 Case study: biomarker discovery pipeline for HCC

A representative workflow for hepatocellular carcinoma (HCC) research involving the *SIGLEC* gene family integrates multiple bioinformatics databases and tools [65]:

(1) Identify the differential expression of *SIGLEC* family genes between HCC and normal tissues using GEPIA [24] (based on TCGA and GTEx data) and validate at the protein level via the Human Protein Atlas (HPA) [66];
(2) Analyze correlations between *SIGLEC* expression and tumor stage/grade using UALCAN [67];

(3) Perform functional enrichment analysis of *SIGLEC*-related genes through WebGestalt [68] and cross-validate with Metascape [69] integrated with DisGeNET [70] for disease associations;
(4) Investigate tumor immune microenvironment interactions by evaluating correlations between *SIGLEC* expression and immune cell infiltration ($CD4^+$ T cells, $CD8^+$ T cells, B cells) as well as immune checkpoint molecules (PD-1, PD-L1, CTLA4) using TIMER [71];
(5) Assess genomic alterations (mutations, copy number variations, mRNA dysregulation) and the survival impact of *SIGLEC* genes in HCC via cBioPortal [10] and Kaplan–Meier Plotter [72] (using TCGA, GEO, and EGA datasets).

This integrated pipeline identifies specific *SIGLEC* family members as potential prognostic biomarkers and regulators of the immune microenvironment in HCC. The databases and tools used can be adapted based on target genes or research objectives.

5 Challenges and future perspectives

The construction of GI cancer databases is an important part of cancer research. The progress and development in cancer omics and related technologies have facilitated the convergence of scientific research and clinical applications, creating publicly accessible repositories of GI cancer data and analytical tools. Cancer omics databases from TCGA, ICGC ARGO, COSMIC, and others are being used to develop new drug targets and diagnostic/prognostic biomarkers [8, 63].

5.1 Gastrointestinal cancer-specific challenges

5.1.1 Disease-specific heterogeneity

GI cancer encompasses diverse subtypes (colorectal, gastric, liver, pancreatic, esophageal) with distinct molecular profiles, posing challenges for database standardization. For example, colorectal cancer has high microsatellite instability (MSI) rates [73], while pancreatic cancer is characterized by *KRAS* mutations and dense stroma [74]. Databases often lack subtype-specific curation, limiting their utility for rare subtypes (e.g., signet ring cell carcinoma). Additionally, liver cancer has distinct etiologies (viral vs. non-viral), requiring stratified data collection that most databases currently lack.

5.1.2 Database limitations

Current databases for GI cancer research are confronted with multiple limitations. On the one hand, many GI cancer-specific databases such as GCGene [48] have not been updated since 2019, resulting in outdated information, and their coverage is uneven, with more resources allocated to colorectal and liver cancer while cholangiocarcinoma and esophageal cancer receive relatively fewer resources. On the other hand, although general databases contain subsets of gastrointestinal cancer data, they are deficient in GI-specific annotations, such as the interactions between the gut microbiome and cancer, and the inconsistencies of data (e.g., in mutation calling pipelines) across different databases further impedes cross-database integration. Additionally, some databases (e.g., the Oncomine database [75]) are plagued by operational difficulties such as funding shortages, which eventually lead to restricted access or even complete termination of operations.

5.1.3 Technical and ethical challenges

GI cancer research also faces notable technical and ethical challenges. In terms of data integration, the heterogeneous nature of related data (encompassing omics, clinical, and microbiome data) calls for sophisticated normalization and integration strategies, which are absent from most existing databases. Regarding privacy protection, GI cancer datasets usually contain sensitive information such as family history of Lynch syndrome, thus requiring the establishment of sound privacy protection frameworks, yet current approaches like the EGA's authorization system [13] are overly cumbersome and have become an obstacle to efficient data sharing. As for artificial intelligence (AI) and machine learning (ML) integration, although AI/ML technologies show great potential for in-depth data mining, their application is hindered by multiple factors, including small sample sizes for rare cancer subtypes, the scarcity of labeled clinical data, and algorithm bias toward populations with adequate representation.

5.1.4 Data governance and FAIR compliance

Many databases lack adherence to the FAIR (Findable, Accessible, Interoperable, Reusable) principles. For example, inconsistent metadata schemas prevent interoperability between CCGD-ESCC and global databases. Application programming interfaces (APIs) are rarely standardized, limiting programmatic access for large-scale analysis. Additionally, data provenance is poorly documented, making it difficult to assess data quality.

5.2 Future directions

5.2.1 Technical innovations

In response to the aforementioned challenges, targeted technical innovations are proposed to advance GI cancer data research and application. For AI/ML integration, it is critical to develop AI-driven tools dedicated to GI cancer data mining—such as deep learning models that leverage GDSC and TCGA data to predict drug sensitivity [76], and natural language processing algorithms that extract biomarkers from unstructured clinical notes in GIDB [49]—while integrating explainable AI to strengthen the credibility of algorithmic predictions for clinical translation. In terms of privacy-preserving data sharing, blockchain technology can be implemented to enable encrypted access to sensitive gastrointestinal cancer data (e.g., germline mutations) while ensuring full data traceability. Regarding multi-omics integration, the development of specialized databases that fuse omics data with gut microbiome, metabolome, and imaging data is recommended.

5.2.2 Database development

To address the limitations of existing resources, targeted strategies for database development are put forward. First, establish subtype-specific databases that focus on rare GI cancer subtypes (e.g., neuroendocrine tumors) as well as etiological subgroups (e.g., viral hepatitis-related liver cancer) to fill the gap of uneven data coverage. Second, implement automated data pipelines to achieve dynamic and real-time database updates—for instance, integrating newly generated TCGA or ICGC ARGO data on a quarterly basis—and set up a warning mechanism for inactive databases while providing corresponding alternative resources. Finally, adhere to the FAIR principles by adopting standardized metadata schemas (e.g., The Global Alliance for Genomics and Health [77]) and interoperable APIs to facilitate cross-database analysis, and develop a specialized

gastrointestinal cancer data portal that aggregates all FAIR-compliant databases for centralized access and utilization.

5.2.3 Collaboration and governance

Effective collaboration and standardized governance frameworks are essential to sustain the advancement of GI cancer data research. First, it is necessary to form global consortia to unify the standards for GI cancer data collection—for example, harmonizing the Chinese population data in CCGD-ESCC with mainstream global datasets—and promote federated analysis strategies to circumvent the risks and limitations of data centralization. Second, multi-party stakeholder engagement should be integrated into database design processes, with clinicians, patients, and data scientists participating jointly to ensure the clinical relevance of database functions, such as incorporating patient input to prioritize quality of life endpoints in survival data statistics. Finally, targeted training and capacity building programs need to be launched, focusing on equipping researchers with the skills for database utilization and AI/ML tool operation, especially for those working in low-resource regions that bear a high burden of GI cancer.

5.2.4 Translational impact

The above technical and database optimization strategies are ultimately aimed at enhancing the translational impact of GI cancer research on clinical practice and drug development. On the clinical application front, databases can be integrated into routine clinical workflows, such as developing a point-of-care tool that leverages ResMarkerDB and GDSC data to deliver personalized treatment recommendations for GI cancer patients. In terms of drug repurposing, cross-database analyses combining resources like GDSC and ClinicalOmicsDB can be conducted to identify potential repurposable drugs for GI cancer, with a notable example being the exploration of metformin for treating *KRAS*-mutant metastatic colorectal cancer [78]. Additionally, integrating real-world evidence derived from electronic health records into dedicated databases enables the validation of clinical trial findings across more diverse GI cancer patient populations, thereby improving the external validity of research outcomes.

6 Conclusion

In conclusion, we introduced several representative GI cancer databases, but due to scope limitations, many other useful resources and tools were not included in the article. These databases can be divided into two categories: General Databases and Specific Databases. This study aims to improve diagnosis, optimize treatment strategies, and effectively prevent cancer by enhancing the understanding of the GI cancer databases. Meanwhile, we discussed the challenges faced in the construction of GI cancer databases and future research directions. We hope that such an introduction can raise awareness and promote the full and effective utilization of these data resources, thereby improving the management of patients with GI cancer.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s12672-026-04489-0.

Additional file1 (XLS 56 kb)

Author contributions
Liu S designed the overall concept and outline of the manuscript; Liu S and Zhang H contributed to the discussion and design of the manuscript; Liu S and Zhang H contributed to the writing and editing of the manuscript, illustrations, and review of the literature. All authors approved the final version of the manuscript.

Funding
This work was supported by the Talent Scientific Research Start-up Foundation of Wannan Medical College (WYRCQD2023045).

Data availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interest
The authors declare no competing interests.

Received: 14 October 2025 / Accepted: 20 January 2026
Published online: 23 January 2026

References

1. Siegel RL, Kratzer TB, Giaquinto AN, et al. Cancer statistics, 2025. CA Cancer J Clin. 2025;75(1):10–45.
2. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
3. Wang S, Zheng R, Li J, et al. Global, regional, and national lifetime risks of developing and dying from gastrointestinal cancers in 185 countries: a population-based systematic analysis of GLOBOCAN. Lancet Gastroenterol Hepatol. 2024;9(3):229–37.
4. Li J, Liu W, Mojumdar K, et al. A protein expression atlas on tissue samples and cell lines from cancer patients provides insights into tumor heterogeneity and dependencies. Nat Cancer. 2024;5(10):1579–95.
5. Zhao JJ, Ong CJ, Srivastava S, et al. Spatially resolved niche and tumor microenvironmental alterations in gastric cancer peritoneal metastases. Gastroenterology. 2024;167(7):1384-1398.e4.
6. Huang KB, Gui CP, Xu YZ, et al. A multi-classifier system integrated by clinico-histology-genomic analysis for predicting recurrence of papillary renal cell carcinoma. Nat Commun. 2024;15(1):6215.
7. Qiu Z, Ji J, Xu Y, et al. Common DNA methylation changes in biliary tract cancers identify subtypes with different immune characteristics and clinical outcomes. BMC Med. 2022;20(1):64.
8. Liu SC, Tang HL, Liu HD, et al. Multi-label learning for the diagnosis of cancer and identification of novel biomarkers with high-throughput omics. Curr Bioinform. 2021;16(2):261–73.
9. Hutter C, Zenklusen JC. The cancer genome atlas: creating lasting value beyond its data. Cell. 2018;173(2):283–5.
10. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
11. Chandrashekar DS, Karthikeyan SK, Korla PK, et al. UALCAN: an update to the integrated cancer data analysis platform. Neoplasia. 2022;25:18–27.
12. Ma L, Zou D, Liu L, et al. Database commons: a catalog of worldwide biological databases. Genomics Proteomics Bioinformatics. 2023;21(5):1054–8.
13. Freeberg MA, Fromont LA, D'Altri T, et al. The European Genome-phenome Archive in 2021. Nucleic Acids Res. 2022;50(D1):D980–7.
14. Raney BJ, Barber GP, Benet-Pages A, et al. The UCSC genome browser database: 2024 update. Nucleic Acids Res. 2024;52(D1):D1082–8.
15. Sondka Z, Dhir NB, Carvalho-Silva D, et al. COSMIC: a curated database of somatic variants and clinical data for cancer. Nucleic Acids Res. 2024;52(D1):D1210–7.
16. Zhang J, Bajari R, Andric D, et al. The international cancer genome consortium data portal. Nat Biotechnol. 2019;37(4):367–9.
17. Solomon BD, Nguyen AD, Bear KA, et al. Clinical genomic database. Proc Natl Acad Sci U S A. 2013;110(24):9851–5.
18. Chang MT, Bhattarai TS, Schram AM, et al. Accelerating discovery of functional mutant alleles in cancer. Cancer Discov. 2018;8(2):174–83.
19. Chang MT, Asthana S, Gao SP, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34(2):155–63.
20. Chakravarty D, Gao J, Phillips SM, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;2017(1):1–16.
21. Suehnholz SP, Nissan MH, Zhang H, et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. Cancer Discov. 2024;14(1):49–65.
22. Zhao H, Yin X, Xu H, et al. LncTarD 2.0: an updated comprehensive database for experimentally-supported functional lncRNA-target regulations in human diseases. Nucleic Acids Res. 2023;51(D1):D199–207.
23. Zhao H, Shi J, Zhang Y, et al. LncTarD: a manually-curated database of experimentally-supported functional lncRNA-target regulations in human diseases. Nucleic Acids Res. 2020;48(D1):D118–26.

24. Tang Z, Li C, Kang B, et al. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45(W1):W98–102.
25. Li C, Tang Z, Zhang W, et al. GEPIA2021: integrating multiple deconvolution-based analysis into GEPIA. Nucleic Acids Res. 2021;49(W1):W242–6.
26. Kang YJ, Pan L, Liu Y, et al. GEPIA3: enhanced drug sensitivity and interaction network analysis for cancer research. Nucleic Acids Res. 2025;53(W1):W283–90.
27. Xing J, Zhai R, Wang C, et al. DiseaseMeth version 3.0: a major expansion and update of the human disease methylation database. Nucleic Acids Res. 2022;50(D1):D1208–15.
28. Lv J, Liu H, Su J, et al., DiseaseMeth: a human disease methylation database. Nucleic Acids Res. 2012;40(Database issue):D1030–D1035.
29. Xiong Y, Wei Y, Gu Y, et al. DiseaseMeth version 2.0: a major expansion and update of the human disease methylation database. Nucleic Acids Res. 2017;45(D1):D888–95.
30. Huang WY, Hsu SD, Huang HY, et al. MethHC: a database of DNA methylation and gene expression in human cancer. Nucleic Acids Res, 2015;43(Database issue):D856–D861.
31. Huang HY, Li J, Tang Y, et al. MethHC 2.0: information repository of DNA methylation and gene expression in human cancer. Nucleic Acids Res. 2021;49(D1):D1268–75.
32. Edwards NJ, Oberti M, Thangudu RR, et al. The CPTAC data portal: a resource for cancer proteomics research. J Proteome Res. 2015;14(6):2707–13.
33. Knol JC, Lyu M, Bottger F, et al. The pan-cancer proteome atlas, a mass spectrometry-based landscape for discovering tumor biology, biomarkers, and therapeutic targets. Cancer Cell. 2025;43(7):1328-1346.e8.
34. Chen J, Liu X, Shen L, et al. CMBD: a manually curated cancer metabolic biomarker knowledge database. Database (Oxford). 2021;2021:baa094.
35. Liu CH, Lai YL, Shen PC, et al. DriverDBv4: a multi-omics integration database for cancer driver gene research. Nucleic Acids Res. 2024;52(D1):D1246–52.
36. Cheng WC, Chung IF, Chen CY, et al. DriverDB: an exome sequencing database for cancer driver gene identification. Nucleic Acids Res, 2014;42(Database issue):D1048–D1054.
37. Chung IF, Chen CY, Su SC, et al. DriverDBv2: a database for human cancer driver gene research. Nucleic Acids Res. 2016;44(D1):D975–9.
38. Liu SH, Shen PC, Chen CY, et al. DriverDBv3: a multi-omics database for cancer driver gene research. Nucleic Acids Res. 2020;48(D1):D863–70.
39. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(Database issue):D955–D961.
40. Alpsoy S, Sezerman OU. Transfer learning with multiomics integration and deep neural networks reveals drug resistance mechanisms in cancer. Sci Rep. 2025;15(1):42295.
41. Moon CI, Elizarraras JM, Lei JT, et al. Clinicalomicsdb: exploring molecular associations of oncology drug responses in clinical trials. Nucleic Acids Res. 2024;52(D1):D1201–9.
42. Peng L, Cheng S, Lin Y, et al. CCGD-ESCC: A Comprehensive Database for Genetic Variants Associated with Esophageal Squamous Cell Carcinoma in Chinese Population. Genomics Proteomics Bioinformatics. 2018;16(4):262–8.
43. Marzec J, Dayem Ullah AZ, Pirro S, et al. The Pancreatic Expression Database: 2018 update. Nucleic Acids Res. 2018;46(D1):D1107–10.
44. Oscanoa J, Ross-Adams H, Ullah A, et al. A central research portal for mining pancreatic clinical and molecular datasets and accessing biobanked samples. Transl Oncol. 2025;62:102550.
45. Kaur H, Bhalla S, Kaur D, et al. CancerLivER: a database of liver cancer gene expression resources and biomarkers. Database (Oxford). 2020;2020:baa012.
46. Kaur H, Dhall A, Kumar R, et al. Identification of platform-independent diagnostic biomarker panel for hepatocellular carcinoma using large-scale transcriptomics data. Front Genet. 2019;10:1306.
47. Kaur H, Bhalla S, Raghava GPS. Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. PLoS ONE. 2019;14(9):e0221476.
48. Zhao M, Chen L, Liu Y, et al. Gcgene: a gene resource for gastric cancer with literature evidence. Oncotarget. 2016;7(23):33983–93.
49. Wang Y, Wang Y, Wang S, et al. GIDB: a knowledge database for the automated curation and multidimensional analysis of molecular signatures in gastrointestinal cancer. Database (Oxford). 2019;2019:baz051.
50. Jiao Z, Zhang X, Xuan Y, et al. Leveraging cfDNA fragmentomic features in a stacked ensemble model for early detection of esophageal squamous cell carcinoma. Cell Rep Med. 2024;5(8):101664.
51. Liu S, Wu J, Xia Q, et al. Finding new cancer epigenetic and genetic biomarkers from cell-free DNA by combining SALP-seq and machine learning. Comput Struct Biotechnol J. 2020;18:1891–903.
52. Shen Y, Wang D, Yuan T, et al. Novel DNA methylation biomarkers in stool and blood for early detection of colorectal cancer and precancerous lesions. Clin Epigenetics. 2023;15(1):26.
53. de Nascimento Lima P, van den Puttelaar R, Knudsen AB, et al. Characteristics of a cost-effective blood test for colorectal cancer screening. J Natl Cancer Inst. 2024;116(10):1612–20.
54. Cai ZR, Zheng YQ, Hu Y, et al. Construction of exosome non-coding RNA feature for non-invasive, early detection of gastric cancer patients by machine learning: a multi-cohort study. Gut. 2025;74(6):884–93.
55. Liu Y, Han J, Kong T, et al. Drivermp enables improved identification of cancer driver genes. Gigascience. 2022;12:giad106.
56. Zhang W, Yang C, Hu Y, et al. Comprehensive analysis of the correlation of the pan-cancer gene HAUS5 with prognosis and immune infiltration in liver cancer. Sci Rep. 2023;13(1):2409.
57. Zhong H, Shi Q, Wen Q, et al. Pan-cancer analysis reveals potential of FAM110A as a prognostic and immunological biomarker in human cancer. Front Immunol. 2023;14:1058627.
58. Guo Q, Zhao L, Yan N, et al. Integrated pan-cancer analysis and experimental verification of the roles of tropomyosin 4 in gastric cancer. Front Immunol. 2023;14:1148056.
59. Liu SC. Comprehensive analysis of clinical and biological value of ING family genes in liver cancer. World J Gastrointest Oncol. 2024;16(6):2592–609.

60. Jin Q, Chang Z, Chen K, et al. Deep learning and multi-omics reveal programmed cell death-associated diagnostic signatures and prognostic biomarkers in gastric cancer. Front Immunol. 2025;16:1690200.
61. Tao Y, Xing S, Zuo S, et al. Cell-free multi-omics analysis reveals potential biomarkers in gastrointestinal cancer patients' blood. Cell Rep Med. 2023;4(11):101281.
62. Yang Y, Lu Y, Jiang W, et al. Individualized prediction of survival benefit from primary tumor resection for patients with unresectable metastatic colorectal cancer. World J Surg Oncol. 2020;18(1):193.
63. Qiu Z, Li H, Zhang Z, et al. A pharmacogenomic landscape in human liver cancers. Cancer Cell. 2019;36(2):179-193.e11.
64. Li M, Zhou T, Han M, et al. cfOmics: a cell-free multi-Omics database for diseases. Nucleic Acids Res. 2024;52(D1):D607–21.
65. Yao S, Chen W, Chen T, et al. A comprehensive computational analysis to explore the importance of SIGLECs in HCC biology. BMC Gastroenterol. 2023;23(1):42.
66. Thul PJ, Akesson L, Wiking M, et al. A subcellular map of the human proteome. Science. 2017;356(6340):eaal3321.
67. Chandrashekar DS, Bashel B, Balasubramanya SAH, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia. 2017;19(8):649–58.
68. Liao Y, Wang J, Jaehnig EJ, et al. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019;47(W1):W199–205.
69. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.
70. Pinero J, Sauch J, Sanz F, et al. The DisGeNET cytoscape app: exploring and visualizing disease genomics data. Comput Struct Biotechnol J. 2021;19:2960–7.
71. Li T, Fan J, Wang B, et al. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. Cancer Res. 2017;77(21):e108–10.
72. Nagy A, Munkacsy G, Gyorffy B. Pancancer survival analysis of cancer hallmark genes. Sci Rep. 2021;11(1):6047.
73. Karri V, Nimkar S, Dalia SM. Adjuvant immunotherapy in microsatellite instability-high colon cancer: a literature review on efficacy, challenges, and future directions. Discov Med. 2025;37(196):808–15.
74. Wang HC, Lin YL, Hsu CC, et al. Pancreatic stellate cells activated by mutant KRAS-mediated PAI-1 upregulation foster pancreatic cancer progression via IL-8. Theranostics. 2019;9(24):7168–83.
75. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004;6(1):1–6.
76. Carli F, Di Chiaro P, Morelli M, et al. Learning and actioning general principles of cancer cell drug sensitivity. Nat Commun. 2025;16(1):1654.
77. Rehm HL, Page AJH, Smith L, et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell Genom. 2021;1(2):100029.
78. Xie J, Xia L, Xiang W, et al. Metformin selectively inhibits metastatic colorectal cancer with the KRAS mutation by intracellular accumulation through silencing MATE1. Proc Natl Acad Sci U S A. 2020;117(23):13012–22.

Publisher's Note