

Reviewed Preprint

v1 • August 10, 2026

Not revised

✉ For correspondence:

liuchaobuaa@buaa.edu.cn

lishenglei@zzu.edu.cn

jhy@zzu.edu.cn

fuyang@zzu.edu.cn

† These authors contributed equally to this work and should be considered as co-first authors.

Competing interests: No competing interests declared

Funding: See [page 18](#)

Reviewing editor: Eyal Itskovits, GlaxoSmithKline, United Kingdom

© 2026, Sun et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

dbGIST: An LLM-Assisted Multi-Omics Resource for Target Exploration and Cross-Dataset Validation in Gastrointestinal Stromal Tumors

Zhao Sun^{1,2,†}, Qian Zhao^{2,†}, Jie-Han Li^{1,†}, Jia-Jing Li^{3,†}, Hao Liu¹, Yan-Xu Guo², Ya-Dong Tang⁴, Fei Yang⁴, Xin Liu⁴, San-Fei Peng¹, Wunan Mi¹, Ge Zhang¹, Zhen Zhang¹, Ming-Liang Yuan¹, Guang-Hui Li¹, Yun-Fei Wang¹, Chao Liu⁵✉, Sheng-Lei Li³✉, Jing-Hua Yang²✉, Yang Fu¹✉

¹Department of Gastrointestinal Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China •

²Clinical Systems Biology Laboratories, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China •

³Department of Pathology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China • ⁴Proteoformics Biotechnologies Inc., Zhengzhou, China • ⁵School of Biological Science and Medical Engineering & School of Engineering Medicine, Beihang University, Beijing, China

eLife Assessment

The paper provides a **valuable**, foundational dataset that will undoubtedly provide substantial value to the cancer research community. The aggregation and harmonization of a broad, rich, multi-omic dataset is **convincing** and will potentially support further downstream research on GIST, which currently has poor supporting datasets. Support for the accompanying claims is more **incomplete**, however: the biological demonstration rests on a single gene tested by one approach, several analyses lack sample sizes and multiple-testing correction, the claims made for the LLM assistant are not backed by direct evaluation, and the curated data matrices would ideally be deposited independently of the web.

<https://doi.org/10.7554/eLife.111894.1.sa3>

Abstract

Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal neoplasms of the gastrointestinal tract, yet GIST-specific omics evidence remains scattered across small cohorts and is not represented as a dedicated disease project in major cancer genomics resources, limiting reproducible target exploration. Here, we present dbGIST (<https://www.dbgist.com> [↗](#)), a dedicated GIST-focused multi-omics resource built to make dispersed GIST evidence searchable, analyzable, and reusable. dbGIST harmonizes data from 37 centers and 1,991 samples, including pathologically verified in-house cohorts, across genomics, bulk transcriptomics, proteomics, phosphoproteomics, and single-cell transcriptomics, and couples these data with curated clinical annotations covering survival, mutation status, risk stratification, metastasis or recurrence, mitotic index, tumor site and size, and imatinib response. The platform supports cohort-level molecular-clinical association, survival, enrichment, immune-infiltration, drug-sensitivity, and single-cell analyses through interactive visualizations, downloadable source data, and public APIs for programmatic access to reusable analysis outputs and visualization-ready data. An optional LLM-assisted interface helps users navigate analyses and interpret outputs. Using MCM7 as a case study, dbGIST linked a resource-derived candidate to survival, risk features, metastatic or recurrent disease, imatinib-response phenotypes, proliferative cell states, and in vitro GIST-cell behavior. dbGIST therefore provides a traceable and interoperable resource for target exploration and precision oncology research in GIST.

1. Introduction

Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal neoplasms of the gastrointestinal tract and are characterized by substantial clinical and molecular heterogeneity [1]. Activating alterations in KIT and PDGFRA have established GIST as a paradigmatic disease for targeted therapy [2, 3]. Nevertheless, primary and secondary resistance to imatinib, metastatic progression, and biologically distinct non-canonical subgroups continue to complicate disease management and therapeutic target identification [4–10]. These challenges underscore the need for systematic resources that enable disease-focused interrogation of candidate genes, pathways, and clinically relevant vulnerabilities in GIST.

Multi-omics profiling has transformed target discovery in cancer [11]. However, GIST remains underrepresented in broad cancer reference resources such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) [12, 13]. Consequently, widely used downstream platforms built on these compendia, including GEPIA2 and TIMER2.0, cannot directly support GIST-specific interrogation [14, 15]. Meanwhile, independent studies have generated valuable GIST datasets spanning genomics, transcriptomics, proteomics, phosphoproteomics, and single-cell transcriptomics, but these resources remain dispersed across repositories, generated by different platforms, and accompanied by uneven clinical annotation [16–18].

Beyond data fragmentation, practical barriers further limit reuse of existing GIST omics data: cross-modal comparison requires substantial manual harmonization, disease-relevant clinical variables are inconsistently organized, and many analytical workflows remain difficult to access for bench scientists and clinicians [19–21]. What the field lacks is therefore not merely another dataset collection, but a practical disease-focused resource for target exploration and cross-dataset validation. At the same time, recent advances in large language models (LLMs) create an opportunity to make such resources more interactive and accessible [22, 23]. For non-bioinformatic users, an LLM-powered conversational layer can provide guided platform navigation, user-oriented question answering, module-specific instructions, and interpretation of analytical outputs generated by the platform, thereby substantially improving usability [24, 25].

To address these needs, we developed dbGIST (<https://www.dbgist.com>), a disease-focused multi-omics resource that integrates dispersed public and three in-house GIST datasets into a unified analytical platform. dbGIST incorporates genomics, transcriptomics, proteomics, phosphoproteomics, and single-cell transcriptomics together with curated clinical annotations, and provides modules for clinical association analysis, survival analysis, functional enrichment, immune-related analysis, drug-response-related exploration, and cell-type-level interrogation. In addition, dbGIST includes an LLM-powered interactive interface that helps users understand how to use the platform, navigate different analytical modules, and interpret the resulting multi-omics outputs [26, 27]. We therefore designed dbGIST not only as a repository, but as an LLM-enhanced practical resource for target exploration and cross-dataset validation in GIST. To demonstrate this utility, we used MCM7 as a representative case and examined its behavior across cohorts, omics layers, and clinically relevant phenotypes, followed by experimental validation in vitro.

2. Results

2.1. Overview of dbGIST

2.1.1 Data and sample type

The dbGIST resource integrates a GIST-specific multi-omics collection spanning genomics, transcriptomics, proteomics, phosphoproteomics, and single-cell transcriptomics from 37 centers and 1,991 samples, including three in-house cohorts (Fig. 1A and 1B). Rather than treating these data as a single pooled matrix, dbGIST organizes them as modality-aware, cohort-level resources with explicit provenance. Public datasets were screened for GIST eligibility, diagnostic support, data type, platform, and clinical annotation availability; the resulting cohort manifests, and inclusion/exclusion records are provided in Supplementary Tables S1–S3. Quantitative outputs underlying key figures are provided as source-data tables where applicable. Transcriptomic and

single-cell datasets were reprocessed from raw files where usable raw data were available, whereas published processed matrices were retained for proteomic/phosphoproteomic studies and other data types where raw reprocessing was not feasible or not applicable. This structure allows users to compare findings across cohorts while preserving the measurement scale and limitations of each source dataset.

The included biospecimens and data objects cover surgical and biopsy-derived GIST specimens, tumor/normal contrasts where available, imatinib-sensitive and resistant settings, metastatic or recurrent disease, and multiple anatomic sites. Clinical variables were manually curated into controlled labels for common GIST-relevant phenotypes, including NIH risk, mutation status, metastatic status, mitotic index, tumor size, imatinib response, and survival endpoints. The three in-house ZZU-GIST cohorts add pathologically confirmed proteomic and single-cell transcriptomic data to the public collection, providing both a resource-building component and an independent context for cross-dataset validation (Fig. 1B [↗](#)).

2.1.2 Home page with LLM-powered chatbot

dbGIST provides a web interface that combines curated multi-omics modules with an optional LLM-assisted interaction layer. The homepage includes a GIST-focused knowledge query interface and optional gene-centric literature support, including external literature-retrieval resources such as OpenScholar, to help users inspect published evidence for queried genes and navigate the resource (Fig. 2A [↗](#); Supplementary Fig. S1A-C [↗](#)). This function is provided only for contextual exploration and user support; it does not contribute to dbGIST cohort-level statistical analyses or numerical results, which are generated by the platform's predefined analytical modules.

Although dbGIST integrates five primary omics layers, its analytical capabilities are organized into six omics-focused sub-pages, with non-coding RNA evidence presented as a separate evidence-aggregation page. The optional LLM-powered interpretation module can be enabled before an omics-specific analysis to support plain-language explanation of figures, module-specific guidance, and follow-up questions (Fig. 2B [↗](#)). This layer is intended to improve accessibility, but it does not alter preprocessing, statistical testing, or returned numerical results. In parallel, the platform provides downloadable visualization outputs and structured result data, and public API endpoints expose selected analysis results and visualization-ready data for external workflow integration.

2.1.3 Omics data analysis page

The dbGIST platform provides eight core analytical modules that enable systematic and interactive exploration of GIST multi-omics data. As summarized in Fig. 1B [↗](#), these modules cover clinical association analysis, driver-gene analysis, functional enrichment, cell-type distribution, immune infiltration, non-coding RNA evidence, phosphoprotein-site analysis, and candidate-drug exploration. Each module operates on curated, precomputed data objects and returns cohort-aware outputs rather than LLM-generated statistics. Users initiate analyses by submitting a gene, after which dbGIST checks data availability and executes the corresponding workflow. Results include publication-ready figures and structured result tables, and selected derived outputs can also be queried through public API endpoints.

2.2. Exploring the multi-omics functions of dbGIST using MCM7 as an illustrative example

To demonstrate how dbGIST can support resource-derived hypothesis generation in gastrointestinal stromal tumor (GIST) research, we selected MCM7 (Minichromosome Maintenance Complex Component 7), a key regulator of DNA replication initiation and cell cycle progression, as a representative query gene [28]. The literature-oriented interface, including external retrieval resources such as OpenScholar, was used only to provide background context for this illustrative query. We then used MCM7 as an illustrative candidate to explore how dbGIST connects transcriptomic expression, clinical associations, single-cell distribution, functional enrichment, immune context, drug-response prediction, and experimental follow-up. No recurrent genomic alteration or quantified phosphoproteomic event for MCM7 was detected in the current GIST

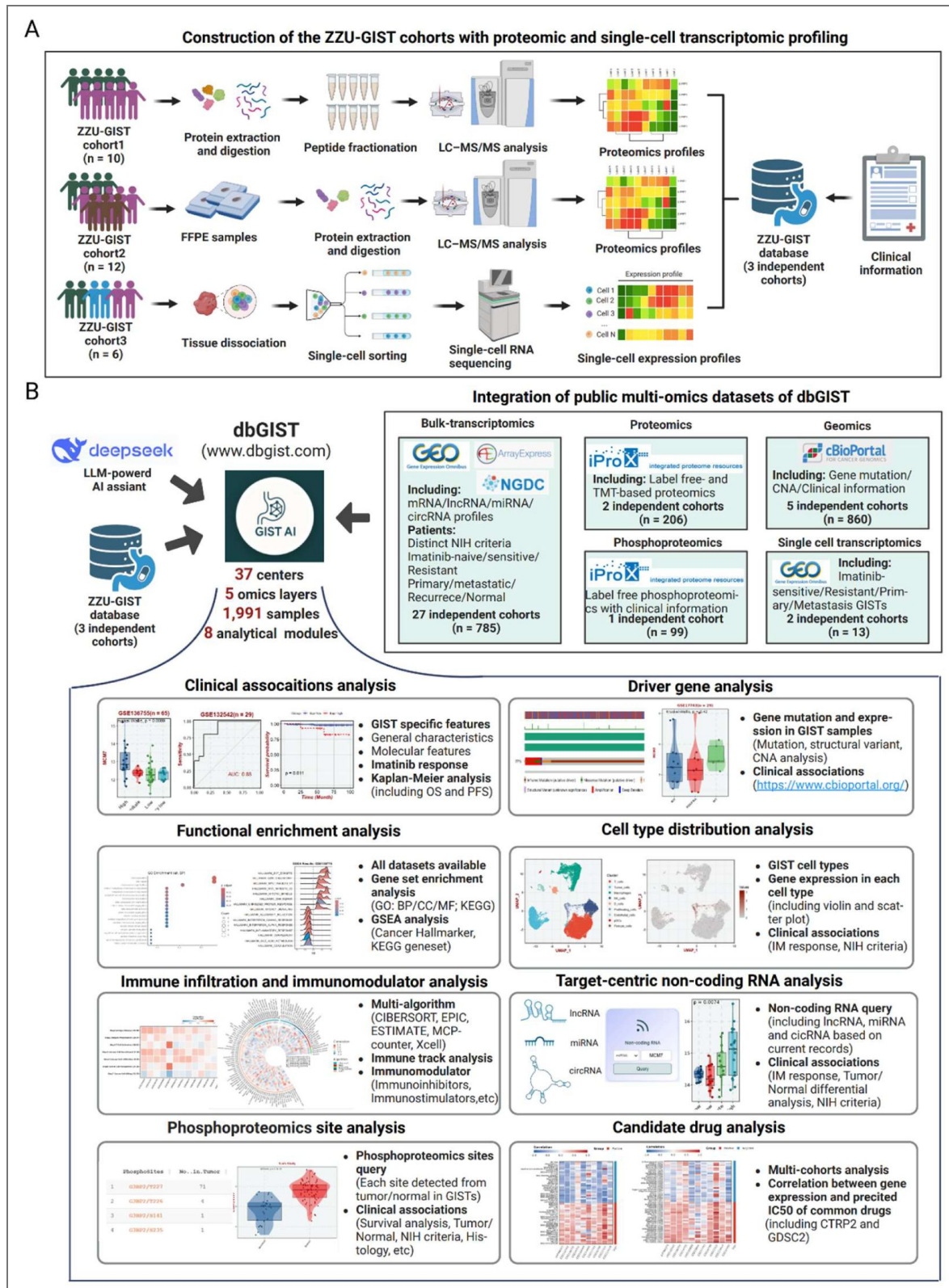


Figure 1. Overview of dbGIST as a GIST-focused multi-omics resource.

(A) Sample processing of in-house GIST proteomics and single-cell transcriptomics. (B) Public and in-house GIST multi-omics data integration and main dbGIST analytical modules.

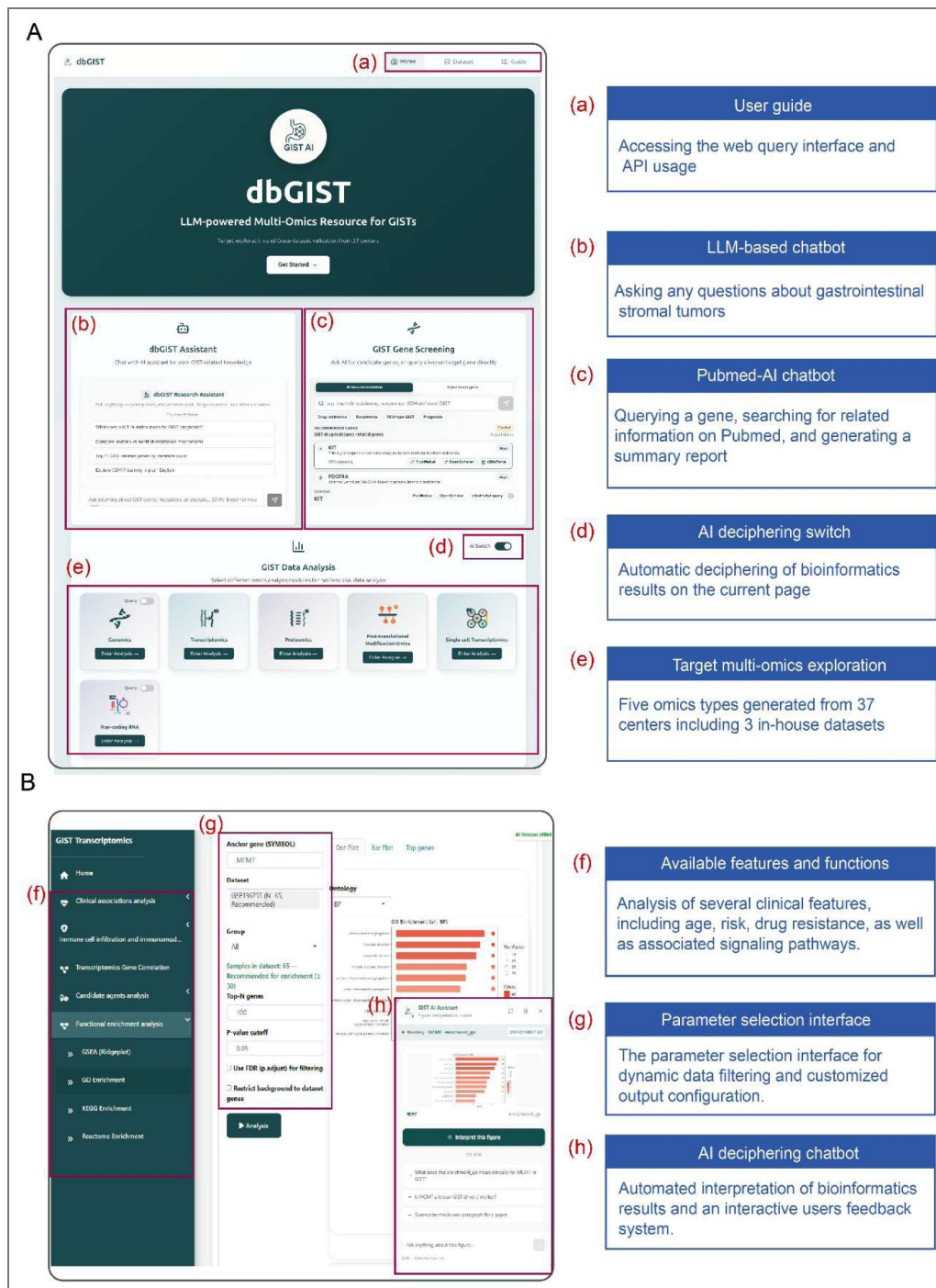


Figure 2. Overview of the dbGIST interface and analytical framework.

(A) Homepage and core functional modules of dbGIST. (a) Top navigation bar providing access to Home, Dataset, and Guide pages for platform navigation and user support. (b) LLM-assisted GIST knowledge chatbot, enabling natural-language queries for biological questions and research guidance. (c) GIST gene screening module for candidate gene querying and target exploration. (d) AI interpretation switch that enables automatic explanation of bioinformatics results on the current page. (e) Entry points for multi-omics data analysis, including genomics, transcriptomics, proteomics, phosphoproteomics, single-cell transcriptomics, and non-coding RNA modules. (B) Representative omics analysis page (transcriptomics module). (f) Left-side navigation panel listing available analytical modules, including clinical association analysis, immune infiltration, gene correlation, and functional enrichment. (g) Parameter configuration interface allowing dataset selection, grouping strategy, and statistical threshold customization for dynamic analysis. (h) AI-assisted interpretation panel that provides automated explanation of analytical results and supports interactive user feedback.

resource; therefore, the case study focuses on data layers in which MCM7 was measured and informative, while the same modules remain available for genes covered by genomic or phosphoproteomic datasets.

2.2.1 Clinical associations

The dbGIST platform enables systematic investigation of associations between molecular features and clinically relevant parameters across integrated omics modalities. Clinical variables were manually curated and harmonized where the source studies allowed, including imatinib treatment response, NIH risk stratification, metastatic status, driver mutation profiles, mitotic index, tumor size, tumor location, age, sex, and survival endpoints. An overview of the primary clinical variables available for each omics dataset is provided in Fig. 3A. When a user initiates a clinical correlation analysis by entering a gene of interest, dbGIST first checks whether the queried feature is available in each applicable cohort and then performs modality-specific association analyses within those cohorts, returning the results for interactive exploration.

As shown in Fig. 3B, proteomic data revealed that high MCM7 expression was significantly associated with poorer overall survival (OS) and progression-free survival (PFS) ($p = 0.011$ and $p = 0.0012$, respectively). Transcriptomic and proteomic analyses further linked MCM7 expression to NIH risk classification (Fig. 3C-D). Elevated MCM7 levels in high-risk patients were observed in three independent transcriptomic cohorts (GSE136755: $p = 0.017$; E-MTAB-373: $p = 0.0081$; GSE47911: $p = 0.038$), a finding corroborated in the in-house ZZU-Cohort1 ($p = 0.024$). MCM7 was also higher in metastatic or recurrent tumors than in non-metastatic cases across three transcriptomic datasets (GSE21315: $p = 0.036$; GSE132542: $p = 6e-06$; E-MTAB-373: $p = 8.7e-07$). Mitotic index showed a positive association with MCM7 expression (Fig. 3F; $p = 1.2e-06$ and $p = 0.032$ in two cohorts), and MCM7 was upregulated in malignant histologic subtypes compared with benign GIST samples ($p = 0.022$). In imatinib-response analyses, MCM7 expression was elevated in resistant cases ($p = 0.00034$) and showed discriminatory performance for resistance (AUC = 0.87; Fig. 3H). In a transcriptomic dataset derived from tumor-associated macrophages isolated from GIST tissues, MCM7 yielded an AUC of 1.0 in this cohort (Fig. 3I). Together, these cohort-level findings nominate MCM7 as a clinically relevant, resource-derived candidate associated with risk, survival, metastatic status, and treatment-response phenotypes.

2.2.2 Cell type distributions

The dbGIST platform enables integrated exploration of single-cell transcriptomic data across multiple independent cohorts. It incorporates three curated single-cell datasets, including two public datasets (GSE162115 and GSE254762) and one in-house ZZU-GIST cohort, processed with standardized quality control, normalization, dimensionality reduction, and cell-type annotation steps (see Methods and Supplementary Methods). Users can query any gene of interest to retrieve its cell-type-specific expression prevalence and perform differential expression analysis across pre-annotated cell populations. Comparative expression analysis can also be performed within clinically stratified cell subsets, such as imatinib-resistant versus sensitive or high-risk versus low-risk groups, enabling cell-level assessment of gene expression patterns in clinical context.

Following standardized processing, the GSE162115, GSE254762, and ZZU-Cohort3 yielded 31,870, 55,429, and 60,437 high-quality cells, respectively. Additional before-QC, after-QC, no-Harmony, and Harmony assessment panels for all three cohorts are provided in Supplementary Figure S4A-L. The cellular composition and subtype distribution for each dataset are summarized in Fig. 4A, 4D, and 4G. MCM7 was consistently enriched in proliferative cell clusters across all three cohorts (Fig. 4B-C, E-F, H-I). These proliferative clusters included malignant cells and actively cycling T cells, both characterized by elevated proliferation markers such as MKI67. dbGIST further linked MCM7 expression in tumor cells to clinically relevant traits, including NIH risk grade, recurrence status, mitotic index, and imatinib resistance (Fig. 4J-M). These results demonstrate that the single-cell module can connect gene-level expression patterns with cell states and clinical annotations across cohorts.

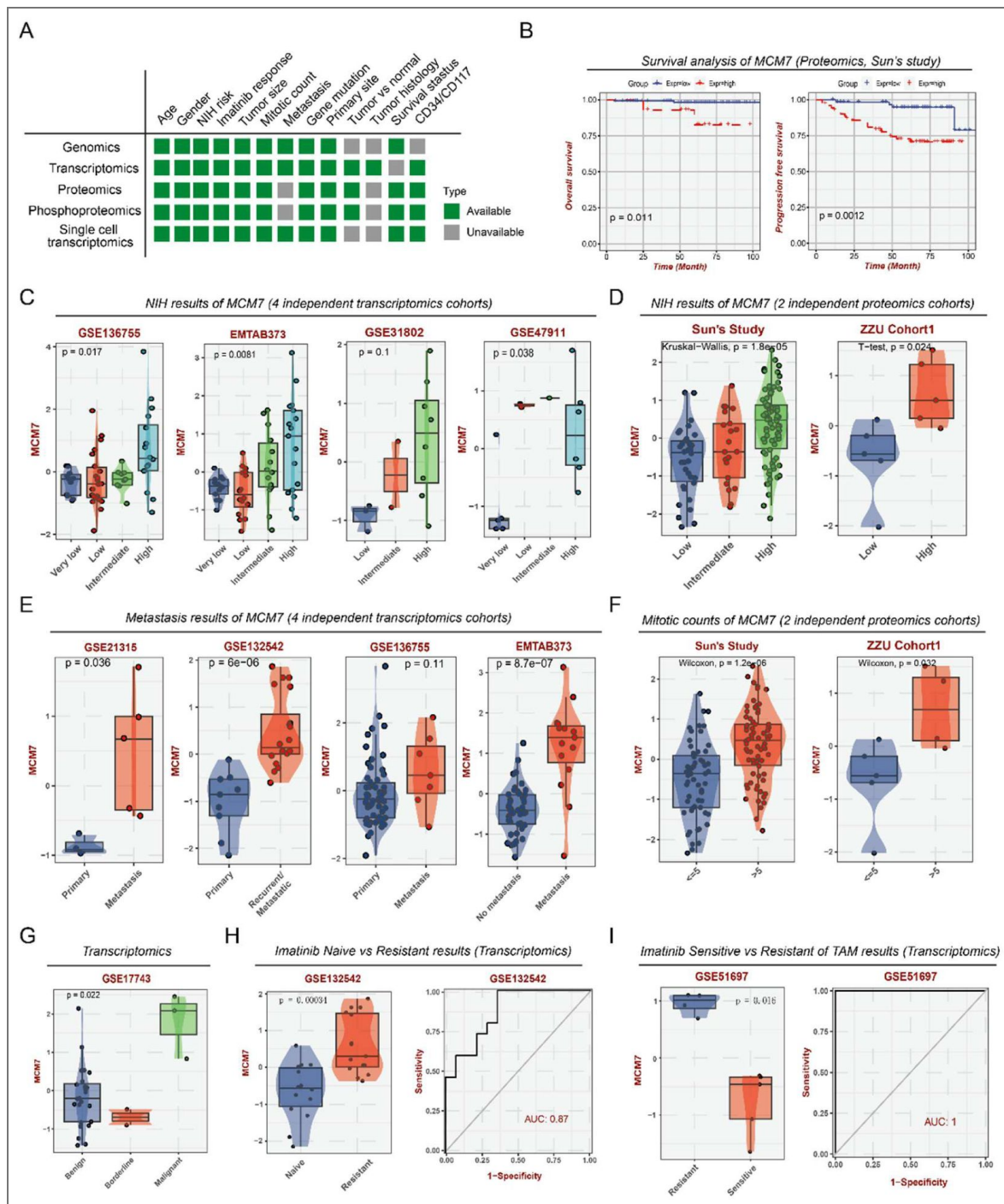


Figure 3. Clinical associations of MCM7 based on integrated omics data.

(A) Summary of clinical annotations available for each omics layer. (B) Kaplan-Meier curves showing overall survival (OS) and progression-free survival (PFS) in the proteomics cohort stratified by MCM7 expression. (C) MCM7 expression across NIH risk groups in four independent transcriptomic cohorts. (D) MCM7 expression across NIH risk groups in two proteomic cohorts. (E) MCM7 expression across primary, metastatic, recurrent/metastatic, and non-metastatic comparison groups in four transcriptomic cohorts. (F) Association between MCM7 expression and mitotic count in two proteomic cohorts. (G) MCM7 expression across histologic categories in a transcriptomic cohort. (H) MCM7 expression by imatinib response and ROC analysis for distinguishing imatinib-resistant from imatinib-naive cases in GSE132542. (I) MCM7 expression by imatinib response and ROC analysis in tumor-associated macrophages from GSE51697. Centre line: median; box bounds: 25th and 75th percentiles.

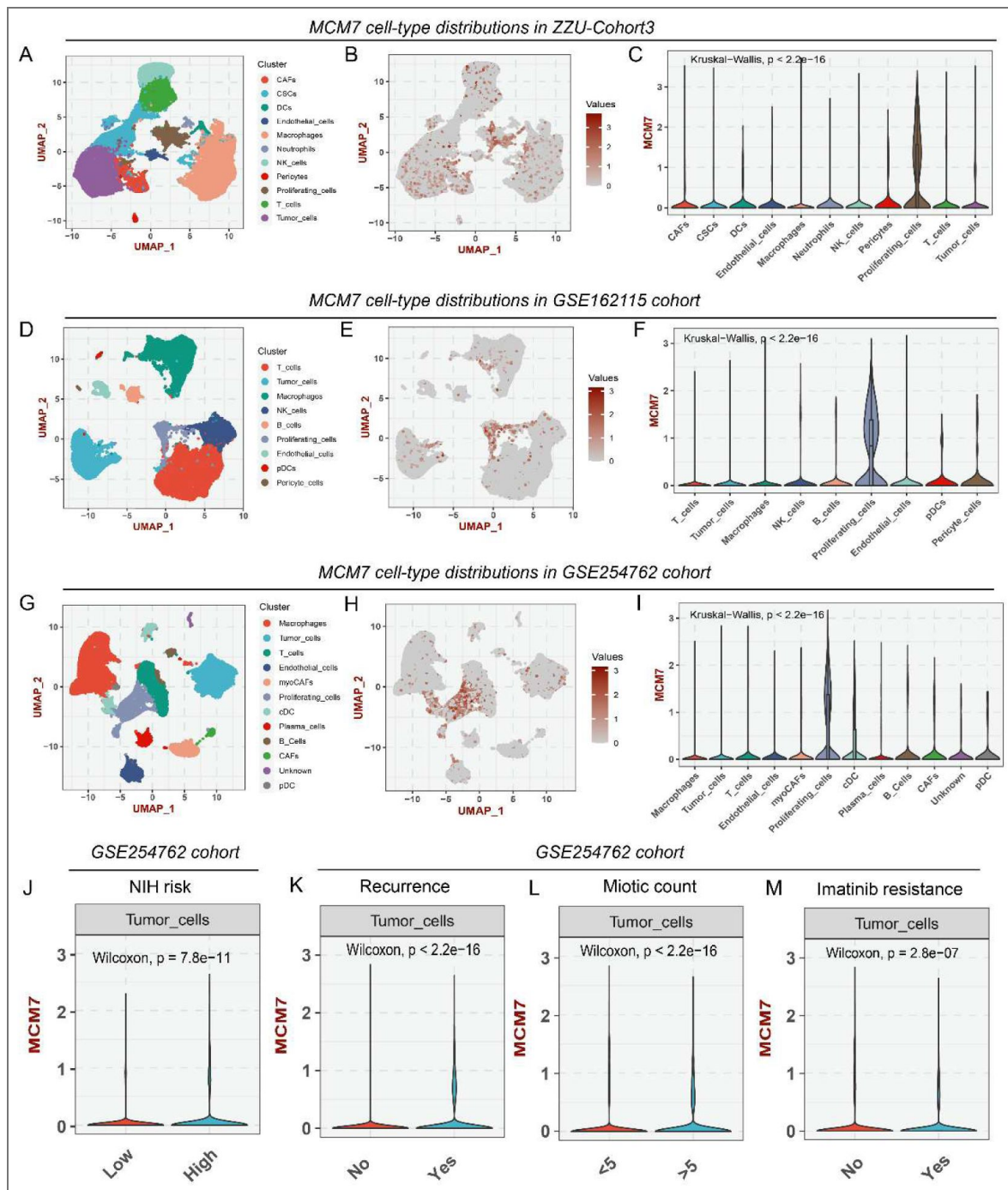


Figure 4. Single-cell distribution and tumor-cell associations of MCM7 in GIST.

(A, D, G) UMAP visualization of annotated cell types in the ZZU-GIST, GSE162115, and GSE254762 cohorts, respectively. (B, E, H) UMAP feature plots showing MCM7 expression in each cohort. (C, F, I) Violin plots showing MCM7 expression across annotated cell types. (J-M) MCM7 expression in tumor cells stratified by NIH risk in GSE162115 and by recurrence, mitotic count, and imatinib resistance in GSE254762. Centre lines indicate medians where shown.

2.2.3 Biological functional enrichment analysis

The dbGIST platform enables users to explore target-gene function through over-representation analysis and gene set enrichment analysis (GSEA), with the integrated workflow shown in Fig. 5A [↗](#). We used this module to characterize the functional landscape associated with MCM7 expression in GIST. Interrogation of the transcriptomic dataset GSE136755 by Hallmark GSEA revealed enrichment of proliferation-associated pathways, including G2M checkpoint and mitotic spindle gene sets in samples with high MCM7 expression (Normalized Enrichment Score [NES] > 2.0, False Discovery Rate [FDR] < 0.001; Fig. 5B [↗](#)). Over-representation analysis of top MCM7-correlated genes further highlighted cell-cycle-related biological processes, including cell division and nuclear division (Fig. 5C [↗](#)), and network visualization illustrated an interconnected mitotic-regulation module (Fig. 5D [↗](#)).

At the protein level, GO biological-process analysis of MCM7-correlated proteins identified DNA replication as the most significantly enriched pathway (Fig. 5E [↗](#)), reinforcing the transcriptomic findings. Network analysis further connected MCM7-associated proteins with cell-cycle and chromosome-organization processes (Fig. 5F [↗](#)). Together, these transcriptomic and proteomic analyses support a reproducible proliferative signature associated with MCM7-high GIST samples, without requiring pooled inference across heterogeneous cohorts.

2.2.4 Immune infiltration, immunomodulators, and candidate compounds

The dbGIST platform also enables systematic investigation of target molecules within the tumor immune microenvironment and candidate-drug context (Fig. 6A [↗](#)). Immune infiltration and tumor microenvironment features are estimated using complementary deconvolution frameworks, including xCell, CIBERSORT, and related tools [29, 30]. Drug-response analyses use pharmacogenomic prediction models trained on large public drug-screening resources, with derived results cached in the platform for rapid retrieval. Users can visualize cohort-level correlations, inspect scatterplots, and download structured result data for downstream interpretation or API-based workflow integration.

As illustrated in Fig. 6B [↗](#), analysis of MCM7 within the cancer-immunity cycle showed no strong association with antigen release or presentation steps. Tumor microenvironment deconvolution further indicated that higher MCM7 expression was inversely correlated with cytotoxic T-cell and natural-killer-cell infiltration (Fig. 6C [↗](#)). dbGIST also identified a positive correlation between MCM7 and MEP (megakaryocyte-erythroid progenitor) cell abundance estimated by deconvolution across independent cohorts (GSE136755: $r = 0.41$, $p = 0.00071$; GSE15966: $r = 0.37$, $p = 0.0065$), with cohort-level scatterplots available for inspection (Fig. 6D [↗](#)). In the drug-sensitivity module, C6-ceramide showed a consistent negative correlation with MCM7 expression across multiple datasets (GSE136755: $r = -0.43$, $p = 0.00044$; GSE15966: $r = -0.54$, $p = 3.9 \times 10^{-5}$; E-MTAB-373: $r = -0.64$, $p = 1.1 \times 10^{-7}$; GSE225819: $r = -0.72$, $p = 5.3 \times 10^{-7}$), nominating it as a candidate compound for further experimental evaluation in MCM7-high GIST contexts (Fig. 6F [↗](#)).

2.2.5 Biological experimental validation of MCM7 in vitro GIST models

To evaluate whether the resource-derived MCM7 signal reflected a testable biological phenotype, we first assessed its relationship to proliferation markers within dbGIST. Across multiple independent patient cohorts, MCM7 showed a positive correlation with the established proliferation marker MKI67 (Ki-67) (Fig. 7A [↗](#)), supporting a link between MCM7 expression and proliferative activity. We then established MCM7-knockdown models in two GIST cell lines, GIST-T1 and GIST-882, using specific siRNAs. Western blot analysis confirmed efficient reduction of MCM7 protein in si-MCM7-transfected cells relative to controls (Fig. 7B, D [↗](#)), and quantitative real-time reverse-transcription PCR further verified reduced MCM7 mRNA expression (Fig. 7C, E [↗](#)).

Functional assays showed that MCM7 knockdown impaired the proliferative capacity of both GIST-T1 and GIST-882 cells (Fig. 7F, I [↗](#)). Wound-healing assays further indicated reduced migratory potential after MCM7 depletion (Fig. 7G, H, J, K [↗](#)). These experiments provide biological support

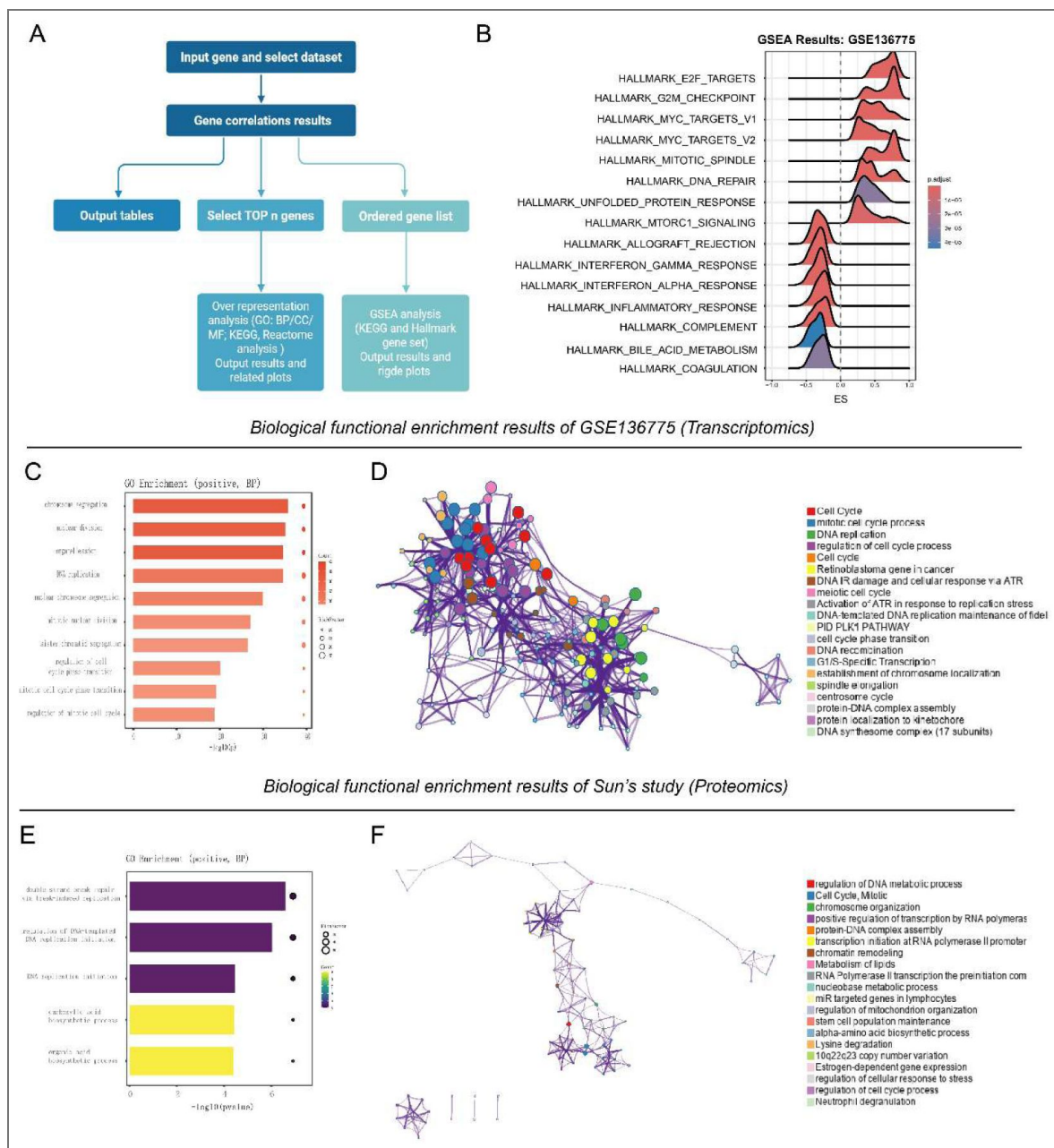


Figure 5. Biological pathway enrichment results of MCM7 in GIST.

(A) Schematic workflow of the integrated bioinformatic analysis pipeline. (B) Gene Set Enrichment Analysis (GSEA) plot of the Hallmark gene set. (C) Bar graph from over-representation analysis (ORA) of the top MCM7-correlated genes. (D) Protein-Protein interaction network of the top MCM7-correlated genes. (E) Bar graph from over-representation analysis (ORA) of the top MCM7-correlated proteins. (F) Protein-Protein interaction network of the top MCM7-correlated proteins.



(A) Flowchart of the integrated bioinformatic pipeline used for tumor immunophenotyping and in silico drug prediction across multiple GIST cohorts. (B) Heatmap showing Pearson correlation coefficients between MCM7 expression and activity of the seven-step cancer-immunity cycle. Red indicates positive correlation and blue indicates negative correlation. (C) Circos plot showing correlations between MCM7 expression and immune cell subtype abundance in the tumor microenvironment, as estimated by multiple deconvolution algorithms. (D) Scatter plots showing positive correlations between MCM7 and MEP (megakaryocyte-erythroid progenitor) cell abundance. (E) Heatmap of correlations between MCM7 expression and predicted drug sensitivity or resistance across analyzed cohorts. (F) Scatter plots showing significant negative correlations between MCM7 expression and predicted response to C6-ceramide.

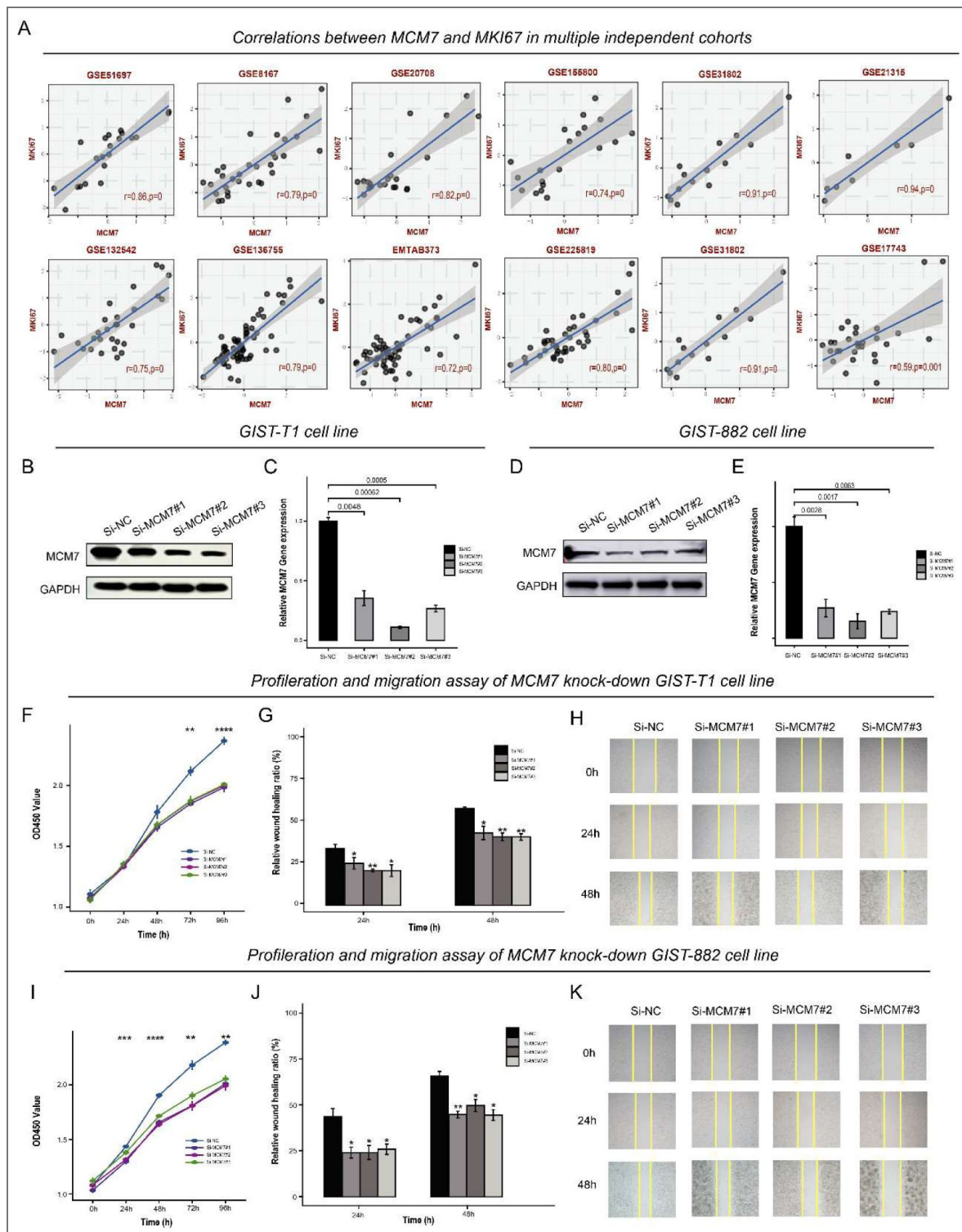


Figure 7. Experimental validation of MCM7 in in vitro GIST models.

(A) Scatter plots showing correlations between MCM7 and MKI67 expression across independent GIST cohorts. (B, D) Western blot validation of siRNA-mediated MCM7 knockdown in GIST-T1 and GIST-882 cells. (C, E) qRT-PCR analysis of MCM7 mRNA expression after siRNA-mediated knockdown in GIST-T1 and GIST-882 cells. (F, I) CCK-8 proliferation assays after MCM7 knockdown in GIST-T1 and GIST-882 cells. (G, J) Quantification of wound-healing assays. (H, K) Representative wound-healing images at 0, 24, and 48 h. Statistical significance: * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

for the dbGIST-derived prioritization of MCM7 and illustrate how the resource can generate experimentally testable target hypotheses.

3. Discussion

In this study, we present dbGIST as a disease-focused multi-omics resource for gastrointestinal stromal tumors rather than a conventional data archive. By integrating datasets from 37 centers, 1,991 samples, three in-house GIST cohorts, and five omics layers, dbGIST addresses a long-standing gap created by the absence of a dedicated GIST cohort in broad cancer genomics consortia. Its value lies in combining curated molecular measurements, clinically meaningful annotations, interactive analytical modules, and accessible interpretation tools within a single GIST-specific research infrastructure.

A central contribution of dbGIST is the curation and harmonization framework underlying the resource. Public GIST omics datasets are heterogeneous in platform, file type, clinical annotation depth, and diagnostic metadata. We therefore verified GIST cohort eligibility by cross-referencing repository records, primary publications, and supplementary clinical files; recorded ICD or ICD-O codes when available; excluded ambiguous mixed tumor collections; and screened TCGA-related categories without identifying a dedicated GIST cohort. Where raw files were usable, transcriptomic and single-cell datasets were reprocessed with modality-appropriate workflows, whereas published processed matrices were retained for data types where raw reprocessing was not feasible or not applicable. Statistical analyses are performed within cohorts, with cross-cohort displays used for comparative interpretation rather than pooled inference. This design makes the scope and limits of harmonization explicit while preserving biological interpretability.

The three in-house ZZU-GIST cohorts further strengthen dbGIST by adding deeply profiled, pathologically confirmed GIST data to the public collection. These cases were confirmed by experienced pathologists using histomorphology and institutional immunohistochemical evaluation, including CD117/KIT and DOG1 where available, with clinicopathological information reviewed from pathology reports and medical records. Importantly, the resource is designed for reuse beyond the web interface. dbGIST provides public API access to analysis results and the result-associated data used for visualization, enabling external tools to query GIST-specific evidence programmatically and incorporate dbGIST outputs into independent analytical workflows.

The MCM7 case study illustrates how dbGIST can support resource-derived hypothesis generation and validation. Across transcriptomic, proteomic, clinical, and single-cell contexts, MCM7 was associated with high-risk features, imatinib response-related phenotypes, adverse survival, and proliferative cell states [28, 31]. Functional perturbation in GIST cell models further supported a role for MCM7 in tumor cell proliferation. We therefore view MCM7 not as an isolated single-gene endpoint, but as a representative example of how dbGIST can connect cross-dataset evidence, multi-omics context, and experimental follow-up for target exploration in GIST.

A cornerstone of dbGIST's utility lies in its LLM-powered framework, which we have successfully implemented to fundamentally transform the user experience by automating the interpretation of analytical results [26]. The conversational AI assistant fundamentally transforms the user experience by automating the interpretation of analytical results, for instance, explaining the clinical implications of a survival curve or the biological context of a pathway enrichment analysis in plain language[32]. More importantly, it provides real-time, dialogue-based troubleshooting, effectively guiding users with limited computational expertise through the analytical workflow. This significantly lowers the barrier for clinicians and wet-lab researchers, empowering them to directly mine the integrated dataset for insights without relying on specialized bioinformatics support. Beyond providing a set of tools, we have established a scalable and transferable framework for building intelligent, domain-specific research platforms (<https://github.com/FuLab-ZhaoSun/dbGIST>). This framework can potentially be adapted to other rare or understudied malignancies that similarly suffer from fragmented data resources and a lack of dedicated analytical services.

Several limitations should be considered. Public GIST datasets remain uneven in sample size, platform coverage, clinical variables, and available raw data, and phosphoproteomic coverage is still more limited than transcriptomic or single-cell coverage. The LLM layer is designed as an optional interpretation and navigation assistant; it does not modify the statistical backend and will require continued maintenance, monitoring, and literature updates. These literature-oriented AI functions, including external retrieval resources such as OpenScholar, are intended to facilitate user navigation and background exploration, but should not be interpreted as systematic reviews or substitutes for manual literature verification [27]. Future versions of dbGIST will expand cohort coverage, improve API documentation, incorporate emerging modalities such as spatial transcriptomics when available, and maintain the platform as a versioned, interoperable, community-oriented resource for GIST research.

4. Methods

4.1 Data collection, cohort eligibility, and pre-processing

We systematically retrieved GIST-related multi-omics datasets from public repositories, including GEO, ArrayExpress, iProX, GSA-Human/NGDC, and cBioPortal, together with in-house cohorts generated at the First Affiliated Hospital of Zhengzhou University. Dataset inclusion required explicit study-level support for a GIST cohort, availability of quantifiable molecular measurements, and accessible clinical annotations. Cohort verification was performed by cross-referencing repository metadata, the primary publication, and supplementary clinical files where available. ICD or ICD-O coding was recorded when explicitly provided, but was not treated as a mandatory inclusion field because such registry-style variables are not consistently exposed in public omics repositories. Datasets labeled only as generic sarcoma, gastrointestinal tumor, or mixed mesenchymal neoplasm were excluded unless a GIST-specific subset could be unambiguously resolved. Detailed cohort manifests, inclusion or exclusion records, and diagnosis-verification notes are provided in [Supplementary Tables S1-S3](#) and [Supplementary Methods](#) Section 1.

In-house ZZU-GIST specimens were obtained from surgically treated patients at the First Affiliated Hospital of Zhengzhou University under institutional ethics approval (Approval No., 2024-KY-1592-001 and 2024-KY-2222-02), with written informed consent obtained from all participants. All included cases were pathologically confirmed as gastrointestinal stromal tumors by experienced pathologists based on histomorphology and institutional immunohistochemical evaluation, including CD117 (KIT) and DOG1 where available. Clinicopathological information was reviewed from pathology reports and medical records, and KIT/PDGFR mutation status was incorporated when available for molecular annotation and subgroup analysis.

Pre-processing was performed in a modality-aware manner rather than through a single merged pipeline. For transcriptomics, raw files were reprocessed with platform-specific workflows whenever usable raw data were available. Genomic cohorts were harmonized to a common mutation-annotation framework centered on the MSK-IMPACT convention. Published quantification matrices were retained for proteomics and phosphoproteomics, whereas single-cell datasets were processed individually with a standardized Seurat-based workflow before being incorporated into the platform as precomputed objects. As part of resource curation, we also screened TCGA soft tissue sarcoma and related disease categories for potential GIST cases, but no dedicated GIST cohort was identified; TCGA was therefore not used as a GIST sample source in dbGIST. Full modality-specific workflows are summarized in [Supplementary Methods](#) Section 1.

Clinical annotations were harmonized through a controlled vocabulary rather than by forcing all cohorts into one pooled analytic matrix. Imatinib response, NIH risk class, mutation category, metastatic status, and other recurrent variables were recoded into standardized labels wherever the source studies allowed. Importantly, inferential analyses in dbGIST are performed within individual datasets, and cross-cohort displays are used for comparative interpretation rather than

for batch-corrected pooled inference. Gene-wise Z-score transformation is therefore used only for visual comparability across cohorts, whereas statistical testing remains cohort-specific. This analytical scope is detailed further in [Supplementary Methods](#) [Section 1](#).

4.2 In-house ZZU-GIST proteomics

The in-house ZZU-GIST proteomics dataset comprised two independent cohorts with distinct sample-processing workflows: liquid nitrogen–frozen tissue specimens in ZZU-Cohort 1 and paraffin-embedded tissue section samples in ZZU-Cohort 2. For ZZU-Cohort 1, proteins were extracted using a RIPA-based workflow, digested with trypsin, fractionated by high-pH reversed-phase chromatography, and analyzed by LC–MS/MS on an EASY-nLC 1200 system coupled to a Q Exactive HF-X Orbitrap mass spectrometer. For ZZU-Cohort 2, deparaffinized tissue sections were subjected to SDS-based protein extraction, tryptic digestion, and LC–MS/MS analysis on an EASY-nLC 1200 system coupled to an Orbitrap Exploris 480 mass spectrometer in data-dependent acquisition mode.

Raw mass spectrometry data were processed using Proteome Discoverer with cohort-specific human reference protein databases and search parameters. False discovery rates were controlled at 1% according to the corresponding database-searching workflows. Protein expression matrices were generated from identified proteins and their label-free quantification intensity values for downstream analyses. Detailed sample-preparation, LC–MS/MS acquisition, database-searching, and quantification procedures are provided in [Supplementary Methods](#) [Section 1.4.3](#).

4.3 In-house ZZU-GIST single-cell transcriptomics

The in-house ZZU-GIST single-cell transcriptomic cohort was generated using GEXSCOPE microfluidic chips for single-cell barcoding, mRNA capture, cDNA synthesis, amplification, and library construction. For the in-house cohort and the two public GIST single-cell cohorts integrated into dbGIST, cells were retained after quality-control filtering based on detected gene number, total UMI count, mitochondrial transcript fraction, and doublet detection where applicable.

Expression matrices were normalized using the LogNormalize method, highly variable genes were selected for dimensionality reduction, and UMAP embeddings were generated from scaled principal-component inputs. Batch-effect assessment was performed by comparing analyses with and without Harmony integration. Before-QC, after-QC, and Harmony-comparison panels for all three scRNA-seq cohorts are shown in [Supplementary Figure S4A–L](#) [Figure S4A–L](#), with detailed preprocessing procedures provided in [Supplementary Methods](#) [Sections 1.4.4 and 2.4](#).

4.4 Online website and AI-assisted implementation

dbGIST is implemented as a modular web resource in which a React and TypeScript front end communicates through a Nginx reverse proxy with a Node.js and Express service layer and multiple independent R Shiny analysis modules. Omics-specific functions are organized as separate yet methodologically aligned applications for genomics, transcriptomics, proteomics, phosphoproteomics, single-cell transcriptomics, and non-coding RNA analysis. The platform also includes an optional AI-assisted interpretation layer that provides narrative explanation and guided query support without altering the underlying statistical results. External literature-retrieval resources, including OpenScholar, are provided only as optional user-support tools for gene-centric literature exploration. They were not used to generate, modify, or validate the omics analyses, statistical tests, survival analysis, enrichment analysis, immune deconvolution, drug-response prediction, experimental validation, or figures reported in this study. All numerical results were generated by predefined dbGIST analytical modules. In other words, the LLM component augments interpretation but does not modify the computational backend. Platform architecture, reproducibility details, and software-version records are summarized in [Supplementary Methods](#) [Sections 3.1–3.3](#) and [Supplementary Table S6](#) [Table S6](#).

4.5 Module-level analyses

Across transcriptomic, proteomic, and phosphoproteomic modules, dbGIST supports cohort-level clinical comparisons, correlation analyses, survival analyses, pathway interpretation, and response-associated modeling using the measurement scale appropriate for each omics layer. Group comparisons are visualized with boxplots or violin plots and tested with paired or non-parametric procedures according to study design and group structure. Survival associations are evaluated by Kaplan-Meier analysis with cutpoint-based dichotomization and log-rank testing, and drug-response associations are summarized either by ROC analysis where labeled response data exist or by pharmacogenomic prediction models where direct response labels are unavailable. Functional interpretation is based on over-representation analysis and gene set enrichment analysis, while immune-context analyses use established deconvolution frameworks applied within cohorts rather than on a pooled matrix. Full statistical settings are summarized in [Supplementary Methods](#) [Section 2](#).

The non-coding RNA module is designed as a read-only evidence-aggregation layer that links user-specified genes to curated miRNA, circRNA, and lncRNA resources with evidence grading and source attribution. Together, these design choices keep the web platform computationally lightweight while preserving methodological transparency and reproducibility. Additional platform-level implementation and supplementary methodological detail are described in [Supplementary Methods](#) [Sections 2 and 3](#).

4.6 Cell culture and transfection

The human gastrointestinal stromal tumor (GIST) cell lines GIST-T1 and GIST-882 were cultured in their respective recommended media supplemented with 10% fetal bovine serum (FBS). Cells were maintained at 37 °C in a humidified atmosphere containing 5% CO₂. For gene silencing experiments, cells were transfected with specific small interfering RNAs (siRNAs) using Lipofectamine 2000 reagent (Life Technologies, USA), strictly adhering to the manufacturer's protocol. Non-targeting siRNAs served as negative controls. All siRNAs were designed and synthesized by Sangon Biotech (China).

4.7 Cell viability assay (CCK-8)

Cell viability was determined using the Cell Counting Kit-8 (CCK-8) assay. Briefly, GIST-T1 and GIST-882 cells, subjected to the respective treatments, were seeded into 96-well plates at a density of 2×10^3 cells per well. At 1, 2, 3, and 4 days after treatment, CCK-8 reagent was added to each well according to the kit's instructions, followed by incubation for a specified duration. The absorbance of each well at 450 nm (OD₄₅₀) was then measured using a microplate reader (DG5032, Hua Dong, Nanjing, China).

4.8 Wound healing assay

The migratory capacity of GIST cells was assessed using a wound healing assay. Transfected GIST-T1 and GIST-882 cells were plated in 6-well plates at a density of 1×10^5 cells per well and allowed to grow until reaching over 90% confluence. A uniform wound was created in the monolayer by scraping with a sterile 200 µL pipette tip. The initial wound width (0 hour) was immediately recorded by capturing images under a microscope. The wound width was measured again after 48 hours of incubation to quantify cell migration.

4.9 Quantitative Real-Time PCR (qRT-PCR)

Total RNA was isolated from cultured cells using Tri-Reagent (TIANGEN BIOTECH, Beijing, China) following the supplier's protocol. RNA concentration and purity were assessed spectrophotometrically. For each sample, 3 µg of total RNA was reverse-transcribed into complementary DNA (cDNA). Quantitative PCR was performed to amplify the genes of interest, with β-actin serving as an internal control for data normalization.

4.10 Western Blot Analysis

Cellular proteins were extracted using RIPA lysis buffer. Protein concentration was quantified with a BCA Protein Assay Kit. Equal amounts of protein (30 µg per lane) were separated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and subsequently transferred onto polyvinylidene fluoride (PVDF) membranes. After blocking with 5% bovine serum albumin (BSA) in TBST for 1 hour at room temperature, the membranes were incubated overnight at 4 °C with primary antibodies against MCM7 (1:1000 dilution) and GAPDH (1:1000 dilution). Following extensive washing, the membranes were incubated with appropriate horseradish peroxidase (HRP)-conjugated secondary antibodies at 37 °C for 1 hour. Protein bands were visualized using an enhanced chemiluminescence (ECL-plus) detection kit, and band intensities were analyzed.

5. Conclusion

In conclusion, dbGIST establishes a GIST-focused multi-omics resource that integrates public and in-house cohorts with diagnostic verification, modality-aware preprocessing, curated clinical annotations, cohort-level analytical modules, and transparent supplementary source data. The platform supports both interactive web-based exploration and programmatic reuse through public APIs that expose selected derived analysis results, while the optional LLM-assisted layer improves navigation and interpretation without modifying the statistical backend. The MCM7 case study demonstrates how dbGIST can connect cross-dataset evidence, multi-omics context, and experimental follow-up to generate testable target hypotheses. By making GIST multi-omics evidence more traceable, reusable, and interoperable, dbGIST provides a practical resource for target exploration and precision oncology research in GIST.

Data availability

dbGIST is available as an online resource through <https://www.dbgist.com>. Dataset-level accession information, cohort manifests, clinical dictionaries, inclusion/exclusion records, and supplementary methodological details are provided in the [Supplementary Methods](#) and [Supplementary Tables](#). Quantitative outputs underlying key figures are provided as source-data tables where applicable. Public datasets used in this study are accessible from GEO (<https://www.ncbi.nlm.nih.gov/geo>), ArrayExpress (<https://www.ebi.ac.uk/biostudies/arrayexpress>), GSA-Human/NGDC (<https://ngdc.cncb.ac.cn/gsa-human/>), iProX (<https://www.iprox.cn>), and cBioPortal (<https://www.cbioportal.org/>). Key code components are hosted on GitHub (<https://github.com/FuLab-ZhaoSun/dbGIST>). dbGIST also provides public API access to selected analysis results and the corresponding visualization-ready data, enabling external tools and workflows to programmatically query derived resource outputs. In-house proteomic and single-cell transcriptomic data were generated under institutional ethics approval 2024-KY-1592-001 and 2024-KY-2222-02 and with written informed consent.

Acknowledgements

This study was supported by Chinese National Natural Science Foundation (92474115 and 82261138558), the Joint Fund of Henan Provincial Research and Development Program for Science and Technology (242301420008), the Scientific Research and Innovation Team of the First Affiliated Hospital of Zhengzhou University (QNCXTD2023022), the Excellent Youth Fund of Henan Natural Science Foundation (212300410075), Zhengzhou University Young Student Basic Research Projects (PhD students) (ZDBJ202518).

Additional information

Author contributions

YF, JY, SL, and CL contributed to the conception and design of the study. ZS, QZ, JH-L, and JJ-L organized the database. ZS, FY, YT, XL, and YW developed the website and performed the statistical analyses. QZ, HL, and YG performed laboratory experiments. ZS, QZ, JH-L, and JJ-L wrote the first draft of the manuscript. SP, GZ, ZZ, MY, WM, and GL wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

Funder	Grant reference number	Author
Chinese National Natural Science Foundation	92474115	Jing-Hua Yang Yang Fu
Chinese National Natural Science Foundation	82261138558	Yang Fu Jing-Hua Yang

Author ORCID iDs

Yang Fu:  <https://orcid.org/0000-0002-8221-3354>

Additional files

[Supplementary Figures](#) 

[Supplementary Methods](#) 

[Supplementary Tables](#) 

References

1. **Blay JY**, et al. (2021) Gastrointestinal stromal tumours. *Nat Rev Dis Primers* **7**:22 <https://doi.org/10.1038/s41572-021-00254-5> | [PubMed](#)
2. **Zhou S**, et al. (2024) KIT mutations and expression: current knowledge and new insights for overcoming IM resistance in GIST. *Cell Commun Signal* **22**:153 <https://doi.org/10.1186/s12964-023-01411-x> | [PubMed](#)
3. **Casali PG**, et al. (2021) Final analysis of the randomized trial on imatinib as an adjuvant in localized gastrointestinal stromal tumors (GIST) from the EORTC Soft Tissue and Bone Sarcoma Group (STBSG), the Australasian Gastro-Intestinal Trials Group (AGITG), UNICANCER, French Sarcoma Group (FSG), Italian Sarcoma Group (ISG), and Spanish Group for Research on Sarcomas (GEIS)(*). *Ann Oncol* **32**:533-541 <https://doi.org/10.1016/j.annonc.2021.01.004> | [PubMed](#)
4. **Heinrich MC**, et al. (2008) Primary and secondary kinase genotypes correlate with the biological and clinical activity of sunitinib in imatinib-resistant gastrointestinal stromal tumor. *J Clin Oncol* **26**:5352-9 <https://doi.org/10.1200/jco.2007.15.7461> | [PubMed](#)
5. **Janeway KA**, et al. (2011) Defects in succinate dehydrogenase in gastrointestinal stromal tumors lacking KIT and PDGFRA mutations. *Proc Natl Acad Sci U S A* **108**:314-8 <https://doi.org/10.1073/pnas.1009199108> | [PubMed](#)
6. **Gao Z**, et al. (2023) N(6)-methyladenosine-modified USP13 induces pro-survival autophagy and imatinib resistance via regulating the stabilization of autophagy-related protein 5 in gastrointestinal stromal tumors. *Cell Death Differ* **30**:544-559 <https://doi.org/10.1038/s41418-022-01107-8> | [PubMed](#)
7. **Zeng X**, et al. (2023) Targeting BCL6 in Gastrointestinal Stromal Tumor Promotes p53-Mediated Apoptosis to Enhance the Antitumor Activity of Imatinib. *Cancer Res* **83**:3624-3635 <https://doi.org/10.1158/0008-5472.can-23-0082> | [PubMed](#)

8. Toulmonde M, et al. (2018) Use of PD-1 Targeting, Macrophage Infiltration, and IDO Pathway Activation in Sarcomas: A Phase 2 Clinical Trial. *JAMA Oncol* **4**:93-97 <https://doi.org/10.1001/jamaoncol.2017.1617> | PubMed
9. Seifert AM, et al. (2017) PD-1/PD-L1 Blockade Enhances T-cell Activity and Antitumor Efficacy of Imatinib in Gastrointestinal Stromal Tumors. *Clin Cancer Res* **23**:454-465 <https://doi.org/10.1158/1078-0432.ccr-16-1163> | PubMed
10. Toulmonde M, et al. (2019) Activity and Safety of Palbociclib in Patients with Advanced Gastrointestinal Stromal Tumors Refractory to Imatinib and Sunitinib: A Biomarker-driven Phase II Study. *Clin Cancer Res* **25**:4611-4615 <https://doi.org/10.1158/1078-0432.ccr-18-3127> | PubMed
11. He X, et al. (2023) Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. *Semin Cancer Biol* **88**:187-200 <https://doi.org/10.1016/j.semcancer.2022.12.009> | PubMed
12. Tomczak K, Czerwińska P, Wiznerowicz M (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol* **19**:A68-77 <https://doi.org/10.5114/wo.2014.47136> | PubMed
13. Zhang J, et al. (2019) The International Cancer Genome Consortium Data Portal. *Nat Biotechnol* **37**:367-369 <https://doi.org/10.1038/s41587-019-0055-9> | PubMed
14. Tang Z, et al. (2019) GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res* **47**:W556-w560 <https://doi.org/10.1093/nar/gkz430> | PubMed
15. Li T, et al. (2020) TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res* **48**:W509-w514 <https://doi.org/10.1093/nar/gkaa407> | PubMed
16. Sun M, et al. (2024) Proteomic Characterization Identifies Clinically Relevant Subgroups of Gastrointestinal Stromal Tumors. *Gastroenterology* **166**:450-465. <https://doi.org/10.1053/j.gastro.2023.11.284> | PubMed
17. Xie F, et al. (2024) Genomic and transcriptomic landscape of human gastrointestinal stromal tumors. *Nat Commun* **15**:9495 <https://doi.org/10.1038/s41467-024-53821-1> | PubMed
18. Liu X, et al. (2024) Deciphering the tumor immune microenvironment of imatinib-resistance in advanced gastrointestinal stromal tumors at single-cell resolution. *Cell Death Dis* **15**:190 <https://doi.org/10.1038/s41419-024-06571-3> | PubMed
19. Wratten L, Wilm A, Göke J (2021) Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods* **18**:1161-1168 <https://doi.org/10.1038/s41592-021-01254-9> | PubMed
20. Attwood TK, et al. (2019) A global perspective on evolving bioinformatics and data science training needs. *Brief Bioinform* **20**:398-404 <https://doi.org/10.1093/bib/bbx100> | PubMed
21. Shi Y, Xu W, Hu P (2025) Out of distribution learning in bioinformatics: advancements and challenges. *Brief Bioinform* **26** <https://doi.org/10.1093/bib/bbaf294> | PubMed
22. Fahrner LJ, et al. (2025) The generative era of medical AI. *Cell* **188**:3648-3660 <https://doi.org/10.1016/j.cell.2025.05.018> | PubMed
23. Zhou J, et al. (2024) An AI Agent for Fully Automated Multi-Omic Analyses. *Adv Sci* **11**:e2407094 <https://doi.org/10.1002/advs.202407094> | PubMed
24. Nalina V, et al. (2025) Advancements in AI for Computational Biology and Bioinformatics: A Comprehensive Review. *Methods Mol Biol* **2952**:87-105 https://doi.org/10.1007/978-1-0716-4690-8_6 | PubMed
25. Guo F, et al. (2025) Foundation models in bioinformatics. *Nat Sci Rev* **12**:nwaf028 <https://doi.org/10.1093/nsr/nwaf028> | PubMed
26. Guo D, et al. (2025) DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**:633-638 <https://doi.org/10.1038/s41586-025-09422-z> | PubMed
27. Asai A, et al. (2026) Synthesizing scientific literature with retrieval-augmented language models. *Nature* **650**:857-863 <https://doi.org/10.1038/s41586-025-10072-4> | PubMed

28. **Stillman B**, Diffley JFX, Iwasa JH (2025) Mechanisms for licensing origins of DNA replication in eukaryotic cells. *Nat Struct Mol Biol* **32**:1143-1153 <https://doi.org/10.1038/s41594-025-01587-5> | PubMed
 29. **Aran D**, Hu Z, Butte AJ (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* **18**:220 <https://doi.org/10.1186/s13059-017-1349-1> | PubMed
 30. **Chen B**, et al. (2018) Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol Biol* **1711**:243-259 https://doi.org/10.1007/978-1-4939-7493-1_12 | PubMed
 31. **Kelly CM**, Gutierrez Sainz L, Chi P (2021) The management of metastatic GIST: current standard and investigational therapeutics. *J Hematol Oncol* **14**:2 <https://doi.org/10.1186/s13045-020-01026-6> | PubMed
 32. **Liu W**, et al. (2025) DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis. *Nat Commun* **16**:2256 <https://doi.org/10.1038/s41467-025-57430-4> | PubMed
- Rink L**, Skorobogatko Y, Kossenkova AV, Belinsky MG, et al. (2009) Gene expression signature predictive of rapid response to imatinib mesylate in Gastrointestinal Stromal Tumor (GIST). NCBI Gene Expression Omnibus. ID 19671739 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15966>

Peer reviews

Reviewer #1 (Public review):

Summary:

This tumour type is missing from the big pan-cancer databases, so none of the popular online analysis tools works for it. That's a real gap, and it's the right one to go after. The authors build an online resource that gathers the scattered public molecular datasets for this disease, adds three of their own patient cohorts, ties everything to clinical data, and exposes interactive tools, downloads, and programmatic access so other people can build on it. To show what it does, they take one gene through the whole platform - clinical, gene-expression, protein, single-cell, immune, and drug-response and then test that gene in cell lines. So there are really two things on offer here: a resource and a practical example of using it. They land very differently.

Strengths:

The resource is the real contribution, and it's done with care. It covers 37 centres and nearly 2,000 samples across five kinds of molecular data, and the authors are honest about provenance: how they screened datasets in or out, where they recorded the diagnostic codes, and why they dropped ambiguous mixed-tumour collections. The key methodological decision is the right one; every analysis runs inside its own cohort, and the cross-cohort views are explicitly "for looking, not for combining." That's exactly how you should treat heterogeneous public data, and they say so plainly instead of quietly pooling everything. Their three pathologist-confirmed cohorts add genuine independent material, so this isn't a re-skin of data that already existed. And because the code and a public access point are actually available, the reuse claim holds.

The example is internally consistent, which is what makes it persuasive. The gene reads higher in higher-risk patients across several independent cohorts and in their own protein data, tracks with the disease spreading and recurring, and lines up with worse survival. The single-cell data put it in the dividing cells; the pathway analysis points to proliferation. Three independent data types landing on the same proliferation story are the strongest part of the biology.

Weaknesses:

The honest problem is that the entire biological story rests on one gene, tested one way. The lab work is two cell lines with the gene knocked down, showing less growth and migration: there is no rescue to confirm the effect is real, no second gene to show the approach generalises, nothing in a living animal. That earns the modest claim: the resource can point you at a candidate worth testing. It does not earn the headline claim that the platform reliably generates good target hypotheses, because we only ever watch it succeed once. One example illustrates a workflow; it doesn't establish a method.

Some of the statistics won't survive scrutiny. The clearest case is a perfect separation between treatment-resistant and treatment-sensitive cases from a single immune cell population, reported with no error bars, no check for information leakage, and apparently from very few samples. A perfect result in that setting is almost always overfitting or a small-sample artefact, not a strong classifier. The same pattern shows up elsewhere: small groups, p-values with no effect sizes or error bars, and no correction for the enormous number of features and cohorts being tested across the whole platform. Separately, one drug result is a correlation against a predicted sensitivity score from a model.

The AI assistant gets far more weight than the evidence supports. Credit where due: the authors are clear and consistent that it only helps interpret and navigate, and never touches the data, the statistics, or the results. That's the correct line to draw, and they hold it. But the assistant itself is never tested, no accuracy numbers, no benchmark, no error analysis, no described way for a human to check what it produces. Calling it something that "fundamentally transforms the user experience" is an assertion, not a finding. And since even the literature feature is admitted not to be a proper systematic review, the prominence of the artificial-intelligence framing runs ahead of what's been shown.

<https://doi.org/10.7554/eLife.111894.1.sa2>

Reviewer #2 (Public review):

Summary:

dbGIST appears to be the first dedicated multi-omics resource worldwide that is specifically focused on GIST.

Strengths:

The main value of the paper is not simply that the authors collected datasets, but that they built a usable resource around them, with cohort-aware analyses, curated clinical labels, interactive visualizations, downloadable results, selected API access, and an optional LLM-assisted interface. The work is solid, and the database is likely to be useful for GIST researchers interested in target discovery, cross-dataset validation, drug-response hypotheses, and translational follow-up.

The MCM7 analysis is a reasonable use case. It shows how a user can start from one candidate gene and then move across transcriptomic, proteomic, clinical, single-cell, immune-related, drug-response, and experimental evidence. I do not see this as the main discovery of the paper, but rather as a practical demonstration of what the database can do. That is appropriate for a resource manuscript.

Weaknesses:

- (1) The authors should make the organization of the platform a little easier to follow. The manuscript refers to five primary omics layers, six omics-focused pages, and eight analytical modules. This structure is understandable after reading the relevant sections, but it may not be immediately obvious to readers. A brief clarification of how the omics layers, web pages, and analytical modules relate to each other would help.
- (2) Since dbGIST is a live web resource, the authors should provide a clear versioning statement. The manuscript should indicate which version of the database corresponds to the analyses and figures reported in the paper, and how future updates will be distinguished from the version evaluated here. This is a small point, but it matters for reproducibility.
- (3) The API function is a strength of the resource, but it is still described rather generally. The authors should give more concrete documentation of what can be accessed through the API, what inputs are required, and what type of output is returned. This could be placed in the supplementary materials. It would make the database more useful for computational users.
- (4) The manuscript should clarify the status of downloadable data. It is clear that figures, source-data tables, and selected derived outputs are available, but it is less clear whether the full processed matrices used internally by the platform are downloadable or only maintained for deployment. This distinction should be stated plainly.
- (5) The statistical reporting in the MCM7 clinical-association analyses needs a little more care. Several p-values are shown across different cohorts and clinical variables. The authors should state whether these are nominal p-values or adjusted p-values. If they are nominal, that is acceptable for a resource demonstration, but the exploratory nature of the analyses should be made clear.
- (6) The ROC analyses for imatinib response should include sample sizes, and confidence intervals for AUC values would be useful if available. Some of the AUC values are high, and without group sizes, it is difficult to judge how stable those estimates are. The authors should avoid implying that these ROC results are validated predictive models.
- (7) The interpretation of MCM7 should be slightly more cautious. MCM7 is a well-known DNA replication and cell-cycle gene, and the single-cell analyses seem to support its association with proliferative cell states. This is biologically consistent, but it also means that MCM7 expression should not be presented as tumour-cell-specific without qualification. The manuscript should frame it mainly as a proliferation-associated signal in the current analysis.
- (8) The drug-response section would benefit from a clearer explanation of the response metric. The authors report correlations between MCM7 expression and predicted response to C6-ceramide, but readers need to know whether the predicted value represents IC50, AUC, sensitivity score, or another metric. The direction of interpretation should also be made explicit, since a negative correlation can mean different things depending on the scoring system.
- (9) The single-cell annotation would be more convincing if the authors provided a compact marker-gene summary for the major cell types in each single-cell cohort. The current description of annotation by source labels, marker inspection, and manual curation is reasonable, but users of the database would benefit from seeing the marker evidence behind the labels.
- (10) The experimental validation section should include a few routine details that are currently not easy to find. The siRNA sequences or target regions, number of biological replicates, statistical tests for the CCK-8 and wound-healing assays, and details of wound-closure quantification should be reported. These additions would make the in vitro part more reproducible.

(11) The wound-healing result should be interpreted with caution. Since MCM7 knockdown reduces proliferation, reduced wound closure could reflect changes in proliferation, migration, or both. Unless proliferation was controlled during the wound-healing assay, the authors should avoid describing this result as purely migratory.

(12) The LLM-related claims should remain conservative. The assistant is a useful feature for navigation, plain-language explanation, and user support, especially for clinicians or wet-lab researchers. However, the strongest statements about the LLM transforming interpretation or automating analysis should be toned down. The important point is that the LLM layer helps users interact with the resource, while the numerical analyses come from predefined dbGIST modules.

<https://doi.org/10.7554/eLife.111894.1.sa1>

Reviewer #3 (Public review):

Summary:

The dbGist dataset/tool would provide substantial value to the cancer research community.

Strengths:

The manuscript presents dbGIST, a dedicated GIST-focused multiomics resource integrating data from 37 centers and ~2k samples across genomics, transcriptomics, proteomics, phosphoproteomics, and single-cell transcriptomics. Given that GIST is virtually absent from major cancer genomics consortia (TCGA, ICGC), this resource fills a genuine gap and represents a valuable contribution to the GIST research community.

(1) The MCM7 case study effectively demonstrates the platform's utility, linking a resource-derived candidate to survival outcomes.

(2) The LLM-assisted interface (dbGIST Assistant) is a reasonable addition for accessibility, lowering the barrier for clinicians and wet-lab researchers, who may not always have the skill set required for proper data analysis, especially for a rich and wide dataset like the dataset in question.

Weaknesses:

(1) Data deposition (major):

While the manuscript references public accessions for raw source datasets and provides a GitHub repository for code, it remains unclear where the **curated, harmonized data matrices** - which represent the core value-add of this work - are independently deposited. Access to these processed data appears to depend entirely on the dbGIST web interface and API. The authors should deposit the harmonized matrices in a persistent, general-purpose repository to ensure long-term availability independent of the web platform.

(2) LLM agent capabilities underspecified:

The manuscript would benefit from a clearer description of the assistant's capabilities and boundaries. Specifically, what tools or actions are available to the LLM agent? Can it execute code against the underlying data, trigger analytical modules programmatically, or is it limited to natural-language explanation of pre-computed results? Clarifying this would help readers assess the scope of the AI layer and distinguish it from agentic platforms that perform computation on behalf of the user.

<https://doi.org/10.7554/eLife.111894.1.sa0>