Journal of Pharmaceutical Analysis 15 (2025) 101065

Contents lists available at ScienceDirect

Journal of Pharmaceutical Analysis

journal homepage: www.elsevier.com/locate/jpa

Original article

A gene set enrichment analysis for cancer hallmarks

Otília Menyhart [a, b, c, 1], William Jayasekara Kothalawala [c, 1], Balázs Győrffy [a, b, c, d, *]

[a] Cancer Biomarker Research Group, Institute of Molecular Life Sciences, Hungarian Research Network, Budapest, 1117, Hungary
[b] National Laboratory for Drug Research and Development, Budapest, 1117, Hungary
[c] Department of Bioinformatics, Semmelweis University, Budapest, 1094, Hungary
[d] Department of Biophysics, Medical School, University of Pecs, Pecs, 7624, Hungary

ARTICLE INFO

Article history:
Received 3 January 2024
Received in revised form
18 June 2024
Accepted 1 August 2024
Available online 3 August 2024

Keywords:
Cancer hallmarks
Candidate hallmark genes
Overrepresentation analysis
Pharmacology
Drug development
New drug targets
Pan-cancer

ABSTRACT

The "hallmarks of cancer" concept provides a valuable framework for understanding fundamental organizing principles common to various cancers. However, without a consensus gene set for cancer hallmarks, data comparison and integration result in diverse biological interpretations across studies. Therefore, we aimed to form a consensus cancer hallmark gene set by merging data from available mapping resources and establishing a framework for mining these gene sets. By consolidating data from seven projects, 6763 genes associated with 10 cancer hallmarks were identified. A cancer hallmarks enrichment analysis was performed for prognostic genes associated with overall survival across 12 types of solid tumors. "Tissue invasion and metastasis" was most prominent in cancers of the stomach ($P = 2.2 \times 10^{-11}$), pancreas ($P = 4.2 \times 10^{-9}$), bladder ($P = 3.3 \times 10^{-8}$), and ovaries ($P = 0.0007$), aligning with their heightened potential to spread. "Sustained angiogenesis" was most prominent in squamous cell carcinomas of the lung ($P = 2.5 \times 10^{-7}$), while "genome instability" showed strong enrichment in lung adenocarcinomas (LUADs) ($P = 1.5 \times 10^{-8}$) and cancers of the liver ($P = 5.5 \times 10^{-10}$), pancreas ($P = 2.1 \times 10^{-5}$), and kidney ($P = 0.018$). Pancreatic cancers displayed the highest enrichment of hallmarks, emphasizing the disease's complexity, while in melanomas and cancers of the liver, prostate, and kidney, a single hallmark was enriched among the prognostic markers of survival. Additionally, an online tool (www.cancerhallmarks.com) that allows the identification of cancer-associated hallmarks from new gene sets was established. In summary, our aim of establishing a consensus list of cancer hallmark genes was achieved. Furthermore, the analysis of survival-associated genes revealed a unique pattern of hallmark enrichment with potential pharmacological implications in different tumor types.

1. Introduction

The lifetime risk of developing cancer in highly developed countries is close to 40%, both in males and females, while the risk of dying from cancer is around 20% [1]. While the term "cancer" comprises a heterogeneous group of diseases with numerous genetic and molecular changes, one unifying feature is the development of abnormal cells that proliferate beyond their natural boundaries. In 2000, Hanahan and Weinberg [2] formulated fundamental organizing principles to provide a framework for understanding features shared by many types of cancer, and they called these features "cancer hallmarks". Initially, six biological capabilities of cancers were presented: sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, sustained angiogenesis, and activating invasion and metastasis. This was updated a decade later with two additional and two emerging hallmarks, respectively: genome instability and mutation, tumor-promoting inflammation, deregulated cellular energetics, and evading immune destruction [3]. In the most recent update from 2022, four additional enabling characteristics were suggested: epigenetic reprogramming, phenotypic plasticity, senescence, and polymorphic microbiomes [4]. The three papers are among the most cited cancer-related publications of the century, indicating the utility of the hallmark approach.

The hallmark concept is helpful for several reasons: it simplifies the intricate nature of cancer phenotypes and genotypes through their consolidation via a preliminary set of fundamental principles, in Hanahan's [4] own words, "helping to distill the complexity of cancer into an increasingly logical science". An understanding of

* Corresponding author. Department of Bioinformatics, Semmelweis University, Budapest, 1094, Hungary.
E-mail address: gyorffy.balazs@yahoo.com (B. Győrffy).
Peer review under responsibility of Xi'an Jiaotong University.
[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.jpha.2024.101065

hallmarks can also aid clinicians in cancer prevention, more accurate diagnosis, and treatment development through the summarizing of functional commonalities [5]. Recognizing interactions across the different hallmark modalities and underlying molecular mechanisms helps clinicians understand the intricacies of malignant transformation and progression [6,7].

Moreover, the adaptability of the hallmark framework facilitates the integration of discoveries as scientific knowledge evolves. For instance, investigating the interactions of RNA molecules with other RNAs, DNAs, and proteins within the context of cancer hallmarks has significantly advanced our understanding of cancer biology and holds promise as a therapeutic strategy [8]. Crosstalk between long noncoding RNAs and micro-RNAs modulates several cancer pathways associated with cancer hallmarks [9]. Furthermore, recent evidence suggests that interactions between the nervous system and developing cancers, at both cellular and molecular levels, can contribute to acquiring hallmark capabilities [10].

In theory, cancer hallmarks and the associated gene sets offer a promising framework for data comparison, integration, and interpretation across studies. High-throughput screening experiments generate extensive gene lists, demanding thorough analysis for biological understanding and prioritization for subsequent investigations. However, despite efforts to define hallmark genes from research data, the lack of a consensus gene set creates challenges in making studies comparable and results in varying biological interpretations [11].

There are different mapping strategies for associating gene sets with hallmarks. For example, researchers can use well-annotated resources, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [12], or conduct a Gene Ontology (GO) enrichment analysis [13] to study the function of hallmark genes and understand their role in specific biological processes. Nevertheless, GO and pathway analyses may produce overlapping or redundant terms that complicate the task of determining precise biological functions. Moreover, different studies employ varying criteria to define their "hallmarks" of cancer. The lack of a standardized reference point may lead to the fragmentation of research findings.

The main aim of this study was to create a unified database of cancer hallmark genes by consolidating recent insights from diverse mapping resources. The list of cancer hallmark genes was compiled from seven different sources. This database serves a twofold function: firstly, it allows for the streamlined exploration of vast gene collections from comprehensive omics studies; secondly, it supports cross-study comparisons, enriching comprehension of cancer biology and revealing overarching patterns in the evolution and progression of malignancies. In serving these functions, the database streamlines the correlation between gene compilations and cancer hallmarks.

The analysis platform, accessible at www.cancerhallmarks.com, allows users to delve into large gene lists within their datasets to detect enrichment based on the 10 hallmarks outlined in the 2000 and 2011 publications on cancer hallmarks [2,3]. The tool identifies the most significant "candidate hallmark genes" associated with cancer in the user's dataset. It provides a downloadable, user-friendly graphical output that visually links genes with biologically relevant functions. Regular updates to the gene list will ensure it stays up-to-date with the latest scientific knowledge.

To illustrate the platform's practicality, we compared the enrichment of cancer hallmarks in gene sets associated with overall survival in 12 major tumor types. The enrichment analysis revealed a distinct "cancer hallmark fingerprint" specific to each tumor type and identified biologically meaningful hallmark enrichment patterns. These findings underscore the utility of the hallmark concept as an effective organizational tool, particularly when establishing a clear connection between genes and biological functions.

2. Methods

2.1. Literature search

A search was performed using The National Center for Biotechnology Information (NCBI) PubMed database. Publications were selected if a list of cancer hallmark genes could be extracted, or if the authors explicitly described the methods for reconstructing a cancer hallmark gene list. Only articles published after the appearance of the second hallmark paper by Hanahan and Weinberg [3], published in 2011, were selected. Only publications that included at least five hallmarks were selected, but most provided information about all 10 (Table S1) [14–20]. Although there were several other publications that included cancer hallmark genes, only seven studies met the requirements with the sufficient detail required to reconstruct associations between genes and cancer hallmarks (Fig. 1).

2.2. Selected publications

The comprehensive cancer hallmark gene set was assembled from seven distinct resources, encompassing six publications [14–19] and the cancer hallmark gene list, which was sourced from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database [20] (Table S1).

The COSMIC database features detailed functional descriptions for each gene [20]. In parallel, a review conducted by Menyhárt et al. [14] included 692 manually curated genes. Four resources utilized GO annotations: Lin et al. [18] accessed cancer hallmark-related GO terms from a previous publication, extracting genes associated with these terms using AmiGo26 [21] (The exact list of GO terms has also been used in two additional studies [22,23]); Iannuccelli et al. [19] integrated the Cancer Census Genes from COSMIC [24] with the gene list obtained from DisGeNET [25], resulting in 91 driver genes linked to seven cancer hallmarks; Knijnenburg et al. [16] merged the curated Pathway Interaction Database (PID), containing human molecular signaling pathways [26], with GO terms [27], deriving 2016 genes associated with 10 hallmarks; and Liang et al. [17] curated genes and pathways from reviews on the effects of carcinogenic chemicals within the Halifax Project [28], subsequently, standardizing the selected genes based on GO terms. Zhang et al. [15] utilized associations between cancer hallmarks and biological pathways: first, hallmark-specific keywords were identified from the literature, and then, based on the collected keywords, candidate hallmark genes were derived from KEGG pathways by the Lucene text-mining software. Genes related to the 10 cancer hallmarks outlined in the 2011, published by Hanahan and Weinberg [3], were thus obtained and merged from the chosen resources to form an integrated cancer hallmark gene set.

2.3. Overrepresentation analysis

The integrated cancer hallmark gene set forms the foundation for conducting enrichment analysis of cancer hallmark genes. In this analysis, known as overrepresentation analysis, specific biological features, such as genes, are examined to determine whether they are more prevalent in a given dataset compared to what would be expected by chance. Using the "enrich" function of the GSEApy Python package, a hypergeometric test is conducted. It compares the frequency of genes in the user's dataset to their frequency in the reference gene set, revealing any significant overrepresentation within the 10 hallmarks of the integrated cancer hallmark gene set [29]. An online web platform was developed using the Flask Python web framework, and operates on an Ubuntu web server. The user's

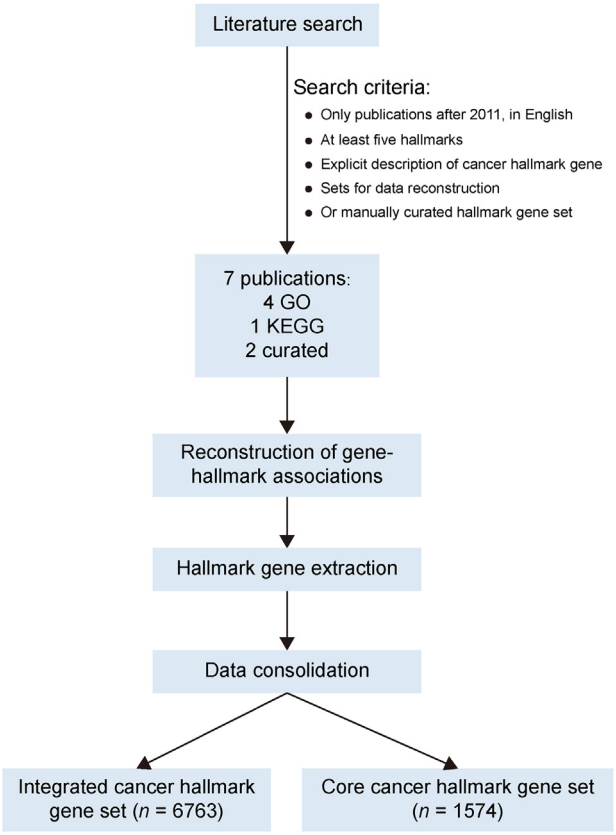

Fig. 1. The data curation and integration process for creating the consolidated cancer hallmark gene sets. GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes.

input gene list is filtered to include only gene symbols approved by the Human Genome Organisation Gene Nomenclature Committee (HGNC), ensuring that only valid gene names are used in the enrichment analysis. Users can access the analysis platform as an open-access online tool using any standard browser, including Firefox, Edge, Chrome, or Safari, at www.carcerhallmarks.com.

2.4. Prognostic gene expression data across 12 tumor types

Clinical follow-up and transcriptomic data for the 12 most common solid tumor types, including breast cancer, colon cancer, prostate cancer, lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), stomach cancer, kidney cancer, bladder cancer, ovarian cancer, pancreatic cancer, liver cancer, and melanoma, were obtained from our previously established pan-cancer database [30], where samples from The Cancer Genome Atlas (TCGA) program had been entered into a single integrated database. Survival analysis was performed using overall survival data for each gene separately. In the Cox regression analysis, each cutoff between the lower and upper quartiles of expression was used, and the false discovery rate was computed to correct for multiple hypothesis testing as described previously [31].

3. Results

3.1. Database of candidate cancer hallmark genes

A database of candidate cancer hallmark genes was created by extracting and merging gene sets from seven resources (Fig. 1 and Table S1). From the selected datasets, genes were identified if they were associated with any of the following 10 cancer hallmarks: sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, sustained angiogenesis, activating invasion and metastasis, genome instability and mutation, tumor-promoting inflammation, deregulated cellular energetics, and evading immune destruction (Fig. 2A). There was a considerable variation in the number of genes within the available resources, ranging between 91 and 7779 (Table S1). The degree of similarity among the seven resources was evaluated using the Tversky index, a similarity coefficient specifically designed for measuring the likeness between datasets. Notably, we found low to moderate similarity between the resources (Fig. S1).

The resulting integrated cancer hallmark gene set consists of 6763 individual genes in total. Most genes were referenced in only one of the seven resources, while 1574 genes were included in at least two gene lists. The option to use either the “integrated cancer hallmark gene set” or the “core cancer hallmark gene set” was added to the online tool (https://cancerhallmarks.com/).

3.2. Genes within individual cancer hallmarks

There was a sizable variation in the number of genes connected to each hallmark. In the integrated cancer hallmark gene set, the number of genes across cancer hallmarks ranged between 574 (“enabling replicative immortality”) and 3574 (“sustaining proliferative signaling”). Note that one gene can be linked to multiple cancer hallmarks. The smallest number of genes was associated with the more recent hallmarks, published in 2011 [3], and the hallmarks “enabling replicative immortality” and “sustaining angiogenesis” (Fig. 2B).

The number of genes in the core cancer hallmark gene set was much lower for each cancer hallmark compared to the integrated cancer hallmark gene set, ranging between 66 (“tumor-promoting inflammation”) and 655 (“sustaining proliferative signaling”). The pattern of fewer genes linked to more recent hallmarks persisted in the core gene set. However, the number of genes linked to the hallmark “activating invasion and metastasis” was disproportionally higher compared to the integrated cancer hallmark gene set (Fig. 2C).

3.3. Overlap across hallmarks

In the integrated cancer hallmark gene set, the most remarkable overlap (2653 common genes) was between the hallmarks “sustaining proliferative signaling” and “evading growth suppressors”. A total of 1314 genes were associated with both “sustaining proliferative signaling” and “tissue invasion and metastasis”, while 1194 genes were associated with both “sustaining proliferative signaling” and “resisting cell death”. The least similar hallmarks were “evading immune destruction” and “genome instability”, with only 46 shared genes (Fig. 3A). For each of the 10 cancer hallmarks, the following 6 genes were identified: *TP53*, *AKT1*, *MAP2K1*, *CTNNB1*, *HRAS*, and *PIK3CB*.

In the core cancer hallmark gene set, the similarity pattern remained similar to that of the integrated cancer hallmark gene set, as “sustaining proliferative signaling” and “evading growth suppressors” remained the most similar, while “evading immune destruction” and “genome instability” remained the least similar cancer hallmarks (Fig. 3B).

3.4. The hallmark enrichment plot

A simple graphical representation was created to illustrate cancer hallmarks that were significantly enriched compared to the

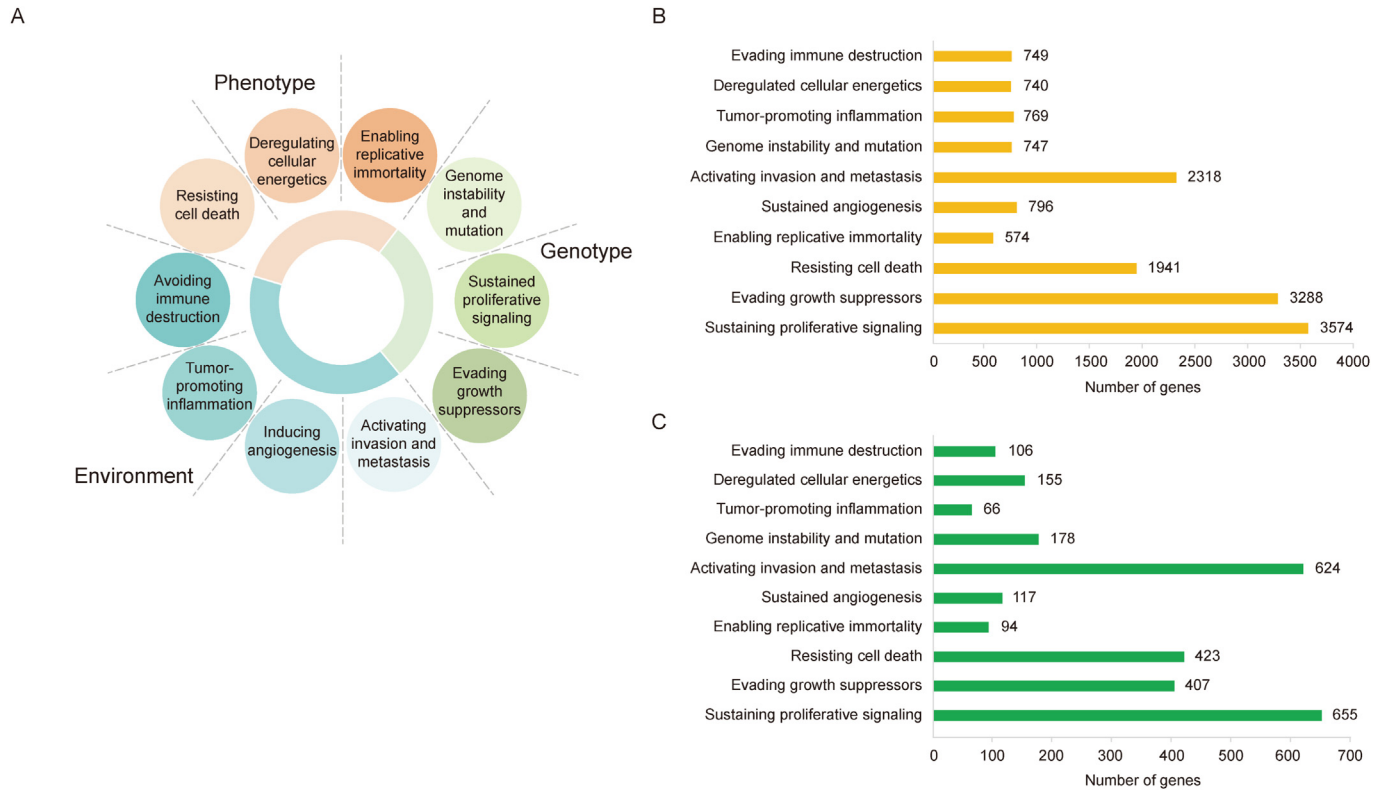

Fig. 2. The 10 cancer hallmarks and the number of corresponding genes. (A) An overview of the 10 cancer hallmarks grouped into three functional classes linked to phenotype, genotype, and tumor environment. (B, C) The number of genes in each of the 10 cancer hallmarks in the integrated cancer hallmark gene set (B) and the core cancer hallmark gene set (C).

reference gene set. The image shows the extent to which the user-submitted gene list aligns with the chosen cancer hallmark gene set (integrated or core) (Fig. 4A).

Each of the 10 cancer hallmarks is represented by differently colored slices, with only the significant ones (adjusted $P < 0.05$) being colored. The size of the slices indicates the level of enrichment compared to the selected reference cancer hallmark gene set. To ensure the best resolution and visual clarity, the axis of the graphical output is adjusted to the smallest *P*-value in the enrichment analysis (see Fig. 4A). If the smallest *P*-value is less than 0.0001, the graph displays three tiers; if less than 0.001, it shows two.

3.5. Distribution of genes across the different hallmarks

An optional analysis allows users to visualize the distribution of genes across different hallmarks. In this feature, black or grey dots represent each of the 10 hallmarks (Fig. 4B). A black dot indicates an overrepresented hallmark compared to the others, while a grey dot signifies an underrepresented hallmark positioned proportionally closer to the inner line. It is worth noting that significantly enriched gene sets can still be underrepresented, and a hallmark with proportionally more genes may not necessarily show significance in the enrichment analysis (see examples in Fig. 5).

Examples of *P*-value-based axis adjustments are also shown in Fig. 5, which includes graphs showing enrichments with two tiers for prostate or kidney cancers, while LUAD, LUSC, and pancreatic cancers display three tiers.

3.6. Analysis of prognostic gene sets

First, all prognostic genes in 12 tumor types were identified. The genes with mean expression cutoff > 100 and $P < 0.05$ were selected for further analysis, and hallmark enrichment analysis was conducted using the gene sets linked to altered prognosis in each of the 12 tumor types separately (Table S2). The conservative core cancer hallmark gene set was employed in the overrepresentation analysis. Genes were classified by their achieved hazard rates (HR) to determine the prognostic relevance: the list of genes correlated to worse ($HR \geq 1.3$) or improved ($HR \leq 0.77$) overall survival outcomes were analyzed separately for each tumor type. These genes are referenced as unfavorable or favorable prognostic genes, respectively, throughout this paper. The number of prognostic genes is summarized in Table S3.

3.6.1. Hallmark enrichment is tissue-specific

By comparing the enrichment of hallmarks in prognostic gene sets, distinct hallmark activities were revealed in each tumor type, forming a unique and tumor-specific cancer hallmark fingerprint. The tumor-specific hallmark enrichments are illustrated in Figs. 5 and 6.

3.6.1.1. Hallmark enrichment of genes linked to unfavorable prognosis. Hallmark enrichment was identified among genes linked to unfavorable prognosis in 10 tumor types. Of these, pancreatic cancer showed the greatest number of enriched hallmarks (8 out of 10), indicating the complexity of the disease and the widespread involvement of diverse biological processes in its development and progression (Fig. 5 and Table S3). Notably, the hallmarks "tissue invasion and metastasis" ($P = 4.2 \times 10^{-9}$), "resisting cell death" ($P = 4.4 \times 10^{-7}$), and "genome instability" ($P = 2.1 \times 10^{-5}$) were particularly prominent in pancreatic cancer (Fig. 5, adjusted *P*-values). These findings suggest that pancreatic cancer, often diagnosed at an advanced stage, is characterized by a heightened potential for spreading to surrounding tissues, aligning with its aggressive and heterogeneous nature.

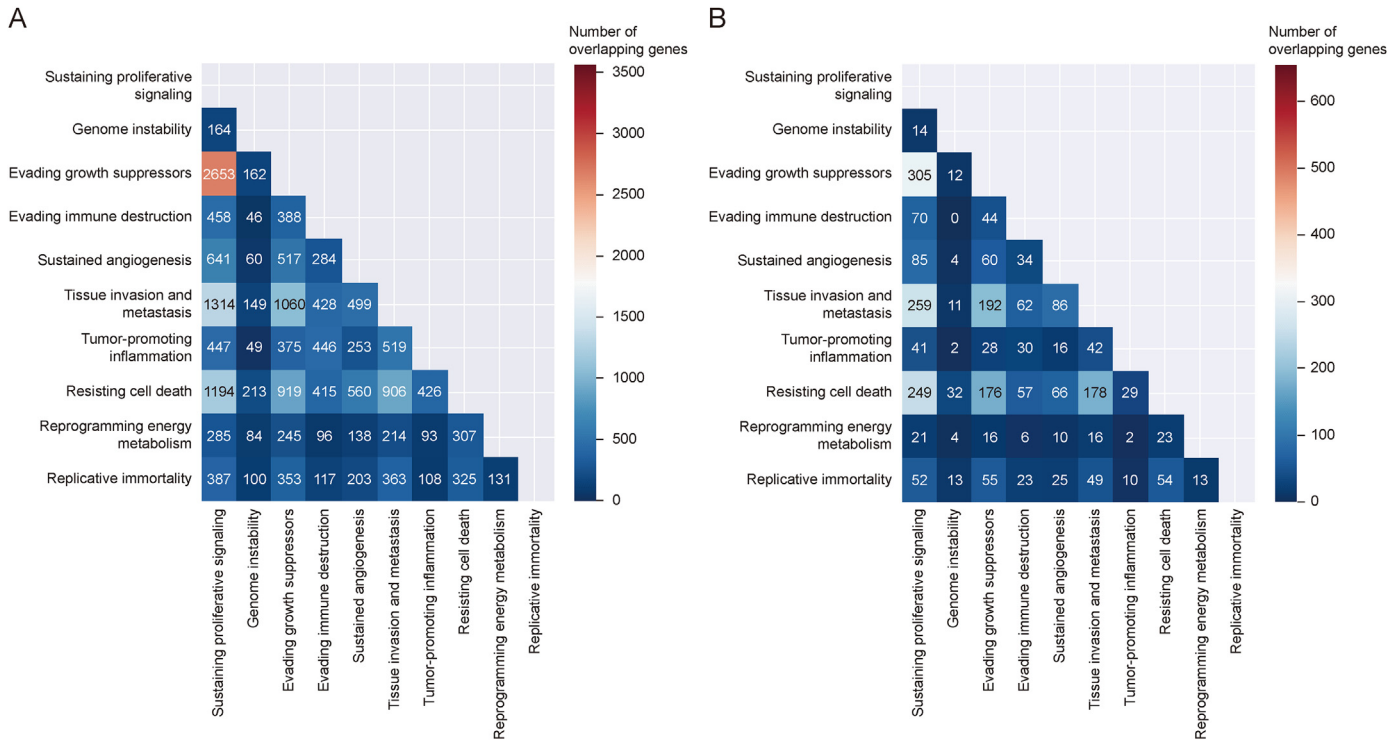

Fig. 3. Similarity matrices across the 10 hallmarks in the (A) integrated and (B) core cancer hallmark gene sets.

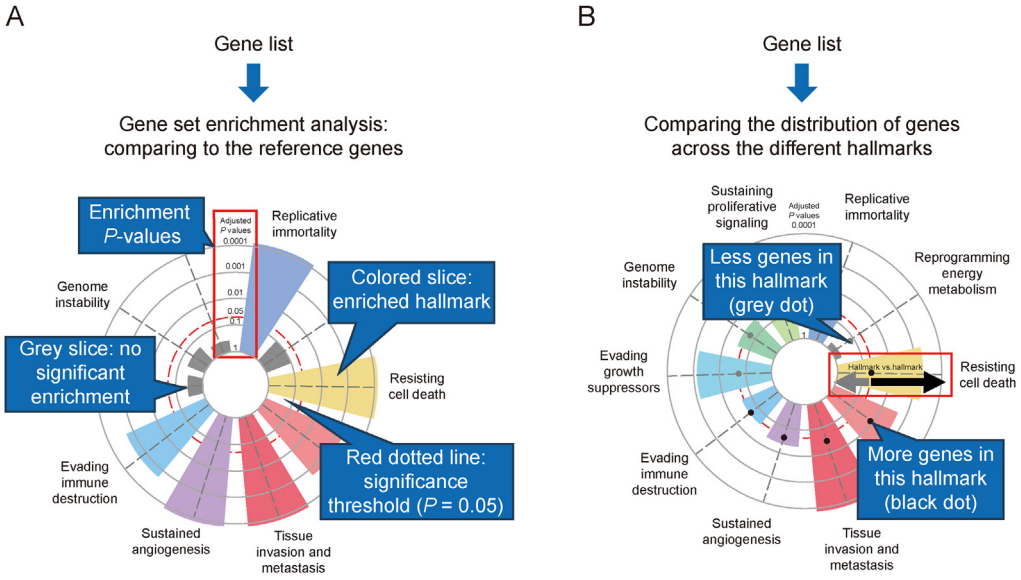

Fig. 4. The hallmark enrichment plot compares the genes we are studying to the set of reference genes and shows how significant this comparison is with colored slices. (A) A red dotted line marks where significance reaches adjusted $P = 0.05$. Only the cancer hallmarks that are significantly enriched are colored. (B) The optional dots inside the colored slices show how genes are spread across different hallmarks compared to each other. Black dots mean there are more genes in that hallmark, while grey dots mean there are fewer. The black and grey arrows compare the distribution of genes across the different hallmarks.

Distinct hallmark enrichment patterns were discerned within lung cancer subtypes. Among the six enriched hallmarks in LUSCs, "sustained angiogenesis" ($P = 2.5 \times 10^{-7}$) and "tissue invasion and metastasis" ($P = 8.9 \times 10^{-7}$) emerged as the most prominent. In contrast, the analysis of LUADs revealed seven enriched hallmarks, with "genome instability" ($P = 1.5 \times 10^{-8}$), "resisting cell death" ($P = 3.3 \times 10^{-5}$), and "reprogramming energy metabolism" ($P = 0.0019$) ranking as the top three.

Similar enrichment patterns were found in stomach and ovarian cancer samples. In stomach cancer, genes related to "tissue invasion and metastasis" ($P = 2.2 \times 10^{-11}$) and "sustained angiogenesis" ($P = 9.2 \times 10^{-11}$) were similarly enriched. In contrast, "tissue invasion and metastasis" ($P = 0.0007$) and "evading growth suppressors" ($P = 0.0038$) were the strongest hallmarks in ovarian cancers. Of the three enriched hallmarks in bladder cancers, "tissue invasion and metastasis" ($P = 3.3 \times 10^{-8}$) was the most prominent.

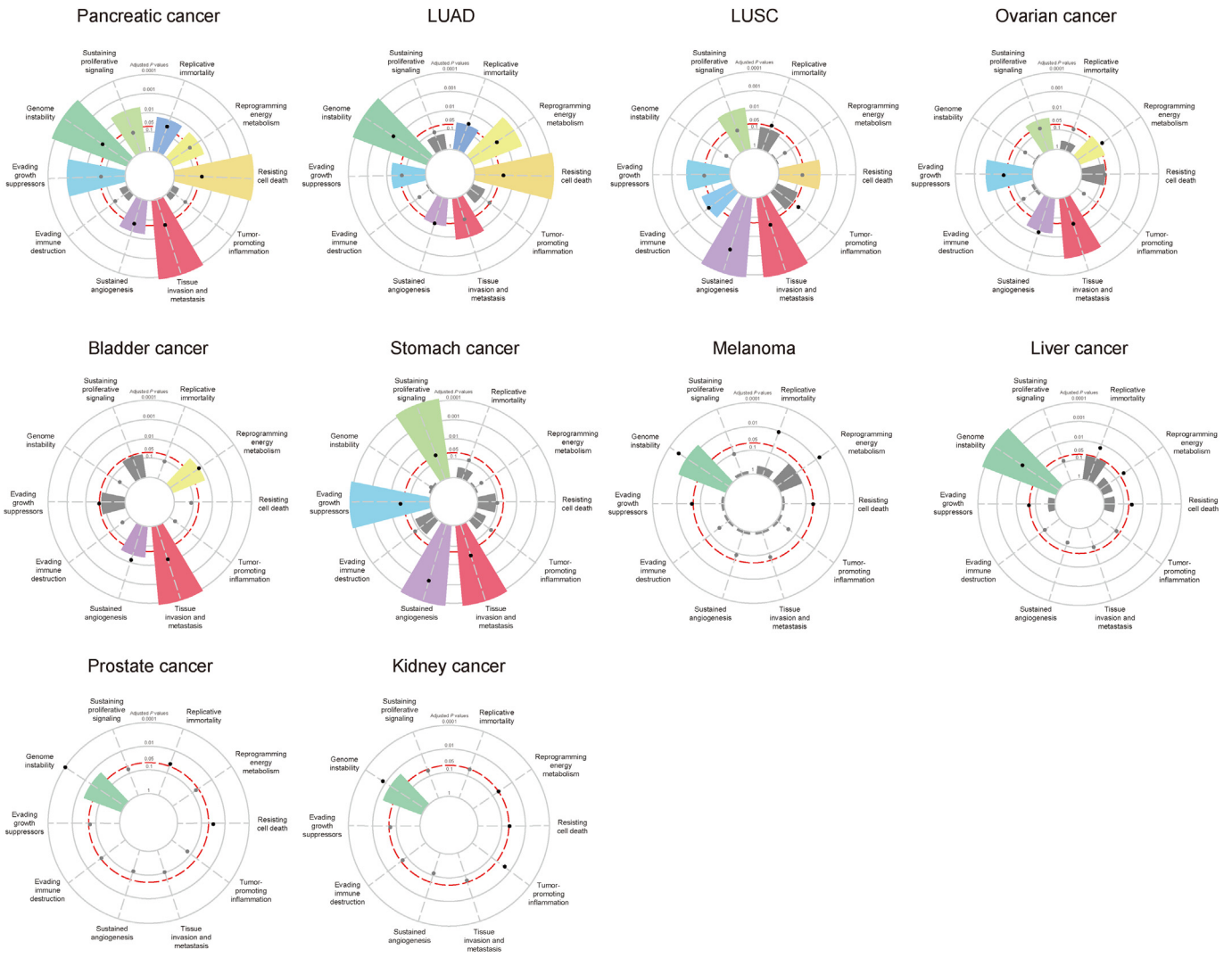

Fig. 5. Cancer hallmark enrichment of genes associated with unfavorable prognosis in 10 tumor types. The red line corresponds to adjusted $P < 0.05$. LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma.

Only a single hallmark, "genome instability" was enriched in melanomas ($P = 0.0076$), and cancers of the liver ($P = 5.5 \times 10^{-10}$), prostate ($P = 0.021$), and kidney ($P = 0.018$) (Fig. 5 and Table S3). Surprisingly, no cancer hallmark enrichment was identified among the unfavorable prognostic genes of colon and breast cancer samples.

In summary, the cancer hallmark "tissue invasion and metastasis" was strongly linked to LUSC, bladder, ovarian, stomach, and pancreas tumors, aligning with their heightened potential to spread. The hallmark "genome instability" was substantially enriched in cancers of the liver, prostate, kidney, and pancreas, as well as in melanoma and LUAD, while "sustained angiogenesis" was substantially enriched in LUSC and in cancers of the stomach and ovary. The hallmark "resisting cell death" was most prominent in LUAD and pancreatic cancer (Fig. 5 and Table S3).

3.6.1.2. Hallmark enrichment of favorable genes. Enrichment analysis on highly expressed genes associated with favorable disease outcomes revealed hallmark enrichment in six distinct tumor groups. These included "evading immune destruction" in melanomas and LUAD, "tumor-promoting inflammation" in melanomas, "genome instability" in colon cancer, stomach cancer, and LUSC, "reprogramming energy metabolism" in colon and kidney cancers, "resisting cell death" in melanomas and cancers of the colon and kidney, "replicative immortality" in colon and kidney cancers, and "sustaining proliferative signaling" in melanomas (Fig. 6 and Table S3).

3.6.2. Central genes linked to multiple cancer hallmarks

The enrichment of cancer hallmarks among genes associated with poor prognosis was observed in 10 tumor types. We found the greatest cancer hallmark enrichments in pancreatic cancer, which exhibited eight significant hallmarks; in LUSC, which exhibited six; and in LUAD, which exhibited seven significant cancer hallmarks (Fig. 5).

Given the intertwined nature of cancer hallmarks, this study focused on central genes associated with multiple hallmarks in these three tumor types. In pancreatic cancer, eight common genes were present in six of the eight enriched cancer hallmarks (Table S4). Among these genes, *MYC* was present in all eight significant hallmarks.

Similarly, in LUSC, 42 genes were present in at least three of the six significantly enriched cancer hallmarks, out of which 10 genes were present in five of the six significantly enriched hallmarks

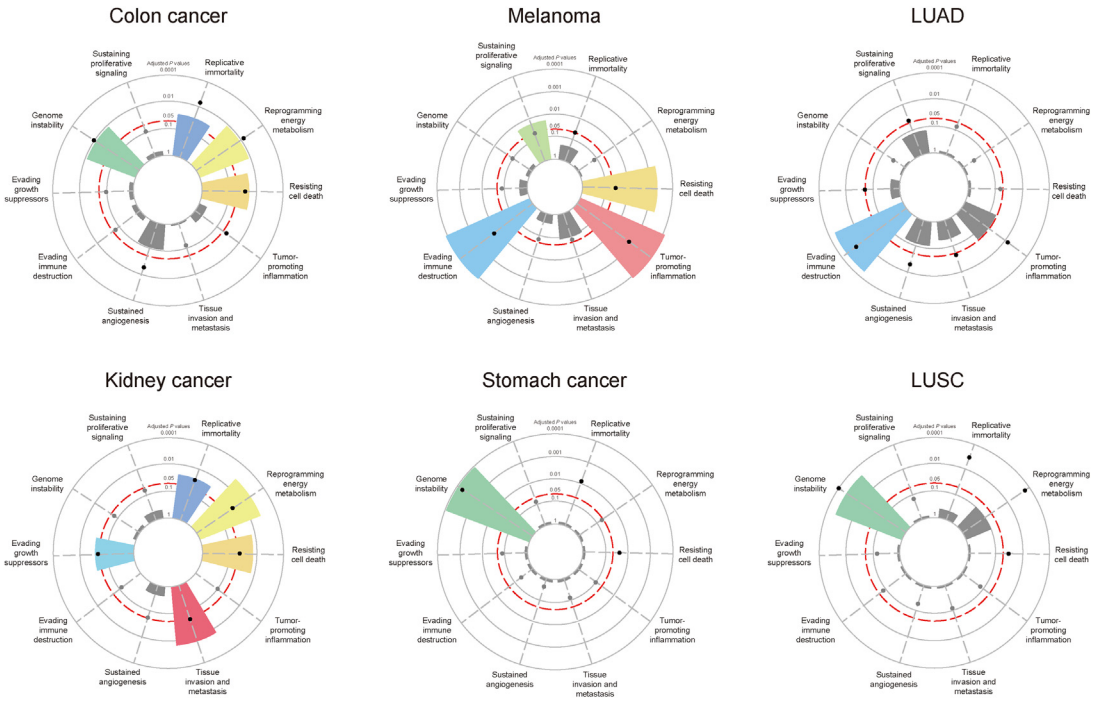

Fig. 6. Cancer hallmark enrichment of genes associated with favorable prognosis in six tumor types. The red line corresponds to adjusted $P < 0.05$. LUAD: lung adenocarcinoma; LUSC: lung squamous carcinoma.

(Table S5). Notably, the *MAPK3* gene was present in all six significantly enriched hallmarks, underscoring its central role.

For LUAD, 38 genes were associated with at least three of the seven significantly enriched hallmarks, out of which nine genes were shared across the cancer hallmarks "evading growth suppressors," "tissue invasion and metastasis", and "resisting cell death" (Table S6). Six well-known cancer-associated genes (*AKT2*, *CCND1*, *EGFR*, *KRAS*, *MAP2K1*, and *PIK3CA*) were shared across five of the seven significantly enriched hallmarks in LUAD (Table S6). The pivotal genes driving multiple cancer hallmarks within each tumor type provide valuable insights into the interconnected nature of tumorigenesis and progression.

3.7. The CancerHallmarks online platform

The online analysis tool established by this project can be accessed at www.cancerhallmarks.com. It can be used to reproduce the statistical results and the hallmark enrichment plots in real-time. In addition, beneath the graphical representation, a curated list of genes linked to each cancer hallmark within the user-submitted data can be displayed. The user can also search for individual genes and discern their association with any identified cancer hallmarks. Another noteworthy feature of the online analysis tool is that users can select between human and mouse reference gene sets. This functionality facilitates the enrichment analysis of cancer hallmark genes in murine research, providing flexibility and relevance in different experimental contexts.

The Molecular Signatures Database (MSigDB) is a widely used and extensive repository of gene sets designed explicitly for conducting gene set enrichment analysis (GSEA) [32]. GSEA is a computational method to determine whether a predefined set of genes exhibits statistically significant, consistent differences between two biological states (e.g., phenotypes). The database combines automated extraction methods with manual curation to create a collection of "hallmark" gene sets. Each hallmark captures specific biological states or processes and exhibits consistent gene expression patterns [33]. Notably, the hallmark gene sets in MSigDB refer to processes within healthy tissues and are not cancer-specific. To harness the data available on the MSigDB platform, the user typically needs to download individual hallmarks and their corresponding gene lists. Subsequent analyses require programming and command-line operations. To streamline and enhance accessibility to this valuable database, we have directly integrated automated MSigDB hallmark enrichment data analysis into our web platform. In addition to the *Integrated* and the *Core hallmark gene sets,* users can paste their gene list of interest into our platform and choose to use the *GSEA-hallmark gene collection* as the reference gene set.

4. Discussion

In the past, the lack of a standardized gene set has impeded data integration across cancer hallmark studies. However, our integrated cancer hallmark database addresses this issue by simplifying the connection between genes and biological functions, offering advantages over traditional GO analysis. Our database offers a standardized reference across molecular and computational biology, genomics, and clinical studies by integrating data on cancer hallmark genes from various sources such as text mining, GO annotations, pathway analysis, and manual curation. The resultant reference gene set promotes a more cohesive understanding of the genetic underpinnings of cancer.

Consolidating data on cancer-hallmark genes from diverse resources offers several advantages. The choice of pathway database can significantly impact enrichment analysis and predictive modeling, as relying solely on GO terms or KEGG pathways may overlook crucial information [34]. A performance comparison of 13 widely used pathway analysis methods suggests varying biases [35]. Conversely, utilizing diverse mapping resources provides

unique information based on different methodologies, resulting in a more comprehensive reference set of hallmark cancer genes.

Integrating information from multiple resources allows for cross-validation of findings, incorporating the latest insights from dynamic sources like text mining while ensuring data reliability through manual curation. This approach strengthens the association of specific genes with cancer hallmarks and mitigates biases inherent in individual resources. Our platform enables users to perform enrichment analysis on 6763 integrated cancer hallmark genes from seven studies. Additionally, users can opt for the core cancer hallmark gene set, derived from genes included in at least two resources, offering flexibility for comprehensive exploration or focused research objectives.

To test the applicability of our concept, we investigated cancer hallmark enrichment among prognostic genes associated with overall survival in 12 cancer types. Cancer hallmark enrichment for unfavorable prognostic genes was identified in 10 cancer types, while cancer hallmark enrichment for favorable prognostic genes was identified in 6 cancer types. The distinct cancer types showed individual enrichment patterns, amounting to a tumor-specific cancer hallmark fingerprint. Pancreatic cancers displayed the highest enrichment of hallmarks, emphasizing the particular complexity of the disease.

Hallmark enrichment was congruent with the molecular subtype of each cancer. Lung cancers, known for their complexity, demonstrate substantial genomic instability, particularly evidenced by a higher genomic instability index (GIS) in metastatic LUAD patients [36]. Notably, a study of LUAD patients observed that lower genomic instability correlated with improved progression-free survival [37]. Our findings align with this pattern, as our study identified enriched genome instability in LUAD among genetic markers associated with unfavorable disease outcomes.

In contrast, in LUSC samples, angiogenesis, a critical phenomenon supporting the progression and metastasis of tumors, was one of the prominently enriched hallmarks among unfavorable prognostic genes. The vascular endothelial growth factor receptor 2 (VEGFR-2) inhibitor, ramucirumab, has been approved for treating both squamous and non-squamous non-small cell lung cancer (NSCLC) and is not associated with an increased risk of respiratory bleeding in squamous histology [38]. Thus, our enrichment analysis might help to identify potential druggability associated with cancer hallmarks.

The hallmark “genome instability” was substantially enriched in melanoma, LUAD, and cancers of the liver, kidney, prostate, and pancreas. Genome instability is mainly attributed to mutations in DNA repair genes, contributing significantly to cancer progression and serving as a crucial prognostic factor [39]. However, genome instability can also stem from abnormal epigenetic modifications, DNA damage originating from telomere issues, and centrosome amplification [40]. The diverse numbers and distinctive patterns of genetic mutations observed in various cancer types indicate distinct tissue and cell-specific mechanisms of tumorigenesis [41]. For instance, melanomas, recognized as highly mutated malignancies, are linked to chronic sun exposure and the detrimental effects of ultraviolet light [42]. Despite these tissue-specific triggers, the shared patterns of cancer hallmark enrichments might suggest a potential common underlying mechanism in tumorigenesis. Emerging evidence supports a connection between lifestyle choices, such as smoking and alcohol consumption, and genomic instability across multiple cancers. Smokers, for example, exhibit elevated total mutational burden when compared to non-smokers in cancers like LUAD, liver, and kidney cancers [43], and cigarette smoking plays an imperative role in the etiology and prognosis of pancreatic cancer as well [44]. Similarly, alcohol consumption has been implicated in genome instability through various mechanisms in liver cancers [45].

In this study, distinctive and mutually exclusive patterns of hallmark enrichment were observed when contrasting unfavorable and favorable prognostic genes within the same tumor types, each with plausible biological explanations. In the case of melanomas and LUAD, the enrichment of immunological hallmarks among highly expressed favorable prognostic genes aligns with the remarkable success observed with immune checkpoint inhibitors (ICI). Notably, nearly 50% of advanced melanoma patients achieve tumor regression and long-term durable cancer control through ICI treatment, a striking contrast to the historical rate of less than 10% [46]. Furthermore, ICIs have demonstrated efficacy in treating NSCLC and colorectal cancer (CRC) as well [47]. Similarly, in colon and stomach cancers, “genome instability” enrichment among highly expressed favorable prognostic genes corresponds with high microsatellite instability's recognized role as a marker of good prognosis. Additionally, it serves as a predictor of heightened sensitivity to immunotherapy-based treatments [48,49].

A notable observation is that cancer hallmarks are not distinct entities but are strongly interlinked. Numerous genes are associated with multiple hallmarks. In our integrated cancer hallmark database, six genes (i.e., *TP53*, *AKT1*, *MAP2K1*, *CTNNB1*, *HRAS*, and *PIK3CB*) were identified as being associated with all 10 investigated cancer hallmarks. These 6 genes are among the most extensively studied in cancer research due to their pivotal roles in tumorigenesis and progression [50–54]. Another well-known example of such central genes are the hypoxia-inducible factors (HIF) that permit cancer cells to adapt to their cellular environment by regulating angiogenesis, invasion, metastasis, and energy metabolism [55]. Such central hub genes may represent crucial nodes in the intricate network of molecular pathways, integrating signals from various upstream regulators and downstream effectors, thereby serving as critical regulators of cancer phenotypes [56].

Our searchable database facilitates the identification of pivotal genes associated with multiple significantly enriched cancer hallmarks, providing valuable insights into the interconnectedness of related processes. In this study, central genes were analyzed in three tumor types, i.e., pancreatic cancers, LUSC, and LUAD, due to their notable enrichment of cancer hallmarks linked with unfavorable prognosis. Notably, *MYC* was present in all eight significantly enriched hallmarks in pancreatic cancer, underscoring the central role played by *MYC*. This finding aligns with existing literature, as *MYC* functions as an essential signaling hub, integrating multiple pathways and driving tumorigenesis and progression in PDAC [57,58]. Additionally, *MAPK3* (*ERK1*), a critical component of the mitogen-activated protein kinase (MAPK) pathway, was shared across all six enriched cancer hallmarks in LUSC, further emphasizing its role in malignant processes. Overexpression of *MAPK3* has been associated with poor survival outcomes in LUSC [59].

Furthermore, a careful look at genes shared across multiple hallmarks within LUSC and LUAD underscores the distinct nature of these two lung cancer subtypes [60]. LUSC-specific genes, notably *EPHA2*, have been associated with the development and progression of squamous cell carcinomas [61]. Meanwhile, LUAD-specific genes such as *EGFR*, *KRAS*, and *MET* play pivotal roles in sustaining proliferative signaling, evading growth suppressors, and facilitating tumor invasion and metastasis [62,63]. Moreover, a divergence in inflammation-related genes was observed between the two. In LUSC, central genes (e.g., *IL1B*, *IL6*, and *CXCL12*) were associated with inflammatory processes, cytokine secretion, and the creation of a pro-tumorigenic microenvironment; by contrast, in LUAD, *TNFRSF1A* and *TRAF2* were identified as central genes integral to inflammatory signaling pathways through their interactions with tumor necrosis factor (TNF) receptors. According to

overall network analysis, critical signaling networks were starkly different between LUSC and LUAD, and even the shared genes were shown to have completely distinct roles in the two diseases [64,65]. The observed dichotomy reflects the unique profiles of LUSC and LUAD, potentially shaping their disease progression trajectories [64,65].

It is essential to acknowledge the inherent limitations in our analysis. The multifaceted nature of tumor characteristics and the intricate pattern of genetic alterations make drawing universal conclusions challenging. Additionally, a primary focus on upregulated genes might overlook valuable insights into how decreased gene expression, especially of tumor suppressors, could inform our understanding of specific cancer hallmarks in tumors. By considering both upregulated and downregulated genes, a more comprehensive understanding of the molecular landscape of cancer might be achieved.

5. Conclusions

This study provides a consensus gene list intricately linked to cancer hallmarks, overcoming previous limitations by facilitating data integration across diverse resources to form a standardized gene set. Through examination of survival-associated genes, unique hallmark enrichment patterns were revealed across different tumor types. Additionally, this study identifies central hallmark genes involved in shared biological processes, a potential area for future drug development efforts. Our findings highlight the significance of the hallmark concept as an organizational tool, aiding understanding of the intricate connections between genes and biological functions.

CRediT authorship contribution statement

Otília Menyhart: Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing, Formal analysis, Visualization. **William Jayasekara Kothalawala:** Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Balázs Győrffy:** Conceptualization, Funding acquisition, Supervision, Visualization, Writing – original draft, Writing – review & editing, Validation.

Declaration of competing interest

The authors declare there are no conflicts of interest.

Acknowledgments

This project was supported by the National Research, Development, and Innovation Office, Hungary (PharmaLab, Grant No.: RRF-2.3.1-21-2022-00015). Otília Menyhart was supported by the Janos Bolyai Scholarship of the Hungarian Academy of Sciences and the Hungarian Scientific Research Fund (Grant No.: OTKA FK147194). The support of ELIXIR Hungary (www.bioinformatics.hu) is acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpha.2024.101065.

References

[1] L. Siegel, K.D. Miller, H.E. Fuchs, et al., Cancer statistics, CA Cancer J. Clin. 71 (2021) 7–33.
[2] D. Hanahan, R.A. Weinberg, The hallmarks of cancer, Cell 100 (2000) 57–70.
[3] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: The next generation, Cell 144 (2011) 646–674.
[4] D. Hanahan, Hallmarks of cancer: New dimensions, Cancer Discov. 12 (2022) 31–46.
[5] L. Yu, Q. Huang, X. Zhou, Identification of cancer hallmarks based on the gene co-expression networks of seven cancers, Front. Genet. 10 (2019), 99.
[6] J. Fares, M.Y. Fares, H.H. Khachfe, et al., Molecular principles of metastasis: A hallmark of cancer revisited, Signal Transduct. Target. Ther. 5 (2020), 28.
[7] K.W. Hon, S.A. Zainal Abidin, I. Othman, et al., The crosstalk between signaling pathways and cancer metabolism in colorectal cancer, Front. Pharmacol. 12 (2021), 768861.
[8] M.M. Gabryelska, S.J. Conn, The RNA interactome in the hallmarks of cancer, Wiley Interdiscip. Rev. RNA 14 (2023), e1786.
[9] R. Bhattacharjee, N. Prabhakar, L. Kumar, et al., Crosstalk between long non-coding RNA and microRNA in cancer, Cell. Oncol. (Dordr.) 46 (2023) 885–908.
[10] D. Hanahan, M. Monje, Cancer hallmarks intersect with neuroscience in the tumor microenvironment, Cancer Cell 41 (2023) 573–580.
[11] Y. Chen, F.J. Verbeek, K. Wolstencroft, Establishing a consensus for the hallmarks of cancer based on gene ontology and pathway annotations, BMC Bioinformatics 22 (2021), 178.
[12] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 28 (2000) 27–30.
[13] S.Y. Rhee, V. Wood, K. Dolinski, et al., Use and misuse of the gene ontology annotations, Nat. Rev. Genet. 9 (2008) 509–515.
[14] O. Menyhárt, H. Harami-Papp, S. Sukumar, et al., Guidelines for the selection of functional assays to evaluate the hallmarks of cancer, Biochim. Biophys. Acta 1866 (2016) 300–319.
[15] D. Zhang, D. Huo, H. Xie, et al., CHG: A systematically integrated database of cancer hallmark genes, Front. Genet. 11 (2020), 29.
[16] T.A. Knijnenburg, T. Bismeijer, L.F.A. Wessels, et al., A multilevel pan-cancer map links gene mutations to cancer hallmarks, Chin, J. Cancer 34 (2015) 439–449.
[17] P.I. Liang, C.C. Wang, H.J. Cheng, et al., Curation of cancer hallmark-based genes and pathways for in silico characterization of chemical carcinogenesis, Database (Oxford) 2020 (2020), baaa045.
[18] K. Lin, L. Song, J. Ma, et al., Identification of cancer hallmark-associated gene and lncRNA cooperative regulation pairs and dictate lncRNA roles in oral squamous cell carcinoma, J. Cell. Mol. Med. 24 (2020) 5213–5223.
[19] M. Iannuccelli, E. Micarelli, P.L. Surdo, et al., CancerGeneNet: Linking driver genes to cancer hallmarks, Nucleic Acids Res. 48 (2020) D416–D421.
[20] J.G. Tate, S. Bamford, H.C. Jubb, et al., COSMIC: The Catalogue Of Somatic Mutations In Cancer, Nucleic Acids Res. 47 (2019) D941–D947.
[21] C.L. Plaisier, M. Pan, N.S. Baliga, A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers, Genome Res. 22 (2012) 2302–2314.
[22] C.L. Plaisier, S. O'Brien, B. Bernard, et al., Causal mechanistic regulatory network for glioblastoma deciphered using systems genetics network analysis, Cell Syst. 3 (2016) 172–186.
[23] V. Thorsson, D.L. Gibbs, S.D. Brown, et al., The immune landscape of cancer, Immunity 48 (2018) 812–830.e14.
[24] Z. Sondka, S. Bamford, C.G. Cole, et al., The COSMIC Cancer Gene Census: Describing genetic dysfunction across all human cancers, Nat. Rev. Cancer 18 (2018) 696–705.
[25] J. Piñero, À. Bravo, N. Queralt-Rosinach, et al., DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res. 45 (2017) D833–D839.
[26] C.F. Schaefer, K. Anthony, S. Krupa, et al., PID: The pathway interaction database, Nucleic Acids Res. 37 (2009) D674–D679.
[27] M. Ashburner, C.A. Ball, J.A. Blake, et al., Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29.
[28] W.H. Goodson 3rd, L. Lowe, D.O. Carpenter, et al., Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: The challenge ahead, Carcinogenesis 36 (2015) S254–S296.
[29] Z. Fang, X. Liu, G. Peltz, GSEApy: A comprehensive package for performing gene set enrichment analysis in Python, Bioinformatics 39 (2023), btac757.
[30] Á. Nagy, G. Munkácsy, B. Győrffy, Pancancer survival analysis of cancer hallmark genes, Sci. Rep. 11 (2021), 6047.
[31] B. Győrffy, Transcriptome-level discovery of survival-associated biomarkers and therapy targets in non-small-cell lung cancer, Br. J. Pharmacol. 181 (2024) 362–374.
[32] A. Subramanian, P. Tamayo, V.K. Mootha, et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. USA 102 (2005) 15545–15550.
[33] A. Liberzon, C. Birger, H. Thorvaldsdóttir, et al., The Molecular Signatures Database (MSigDB) hallmark gene set collection, Cell Syst. 1 (2015) 417–425.
[34] S. Mubeen, C.T. Hoyt, A. Gemünd, et al., The impact of pathway database choice on statistical enrichment analysis and predictive modeling, Front. Genet. 10 (2019), 1203.
[35] T.M. Nguyen, A. Shafi, T. Nguyen, et al., Identifying significantly impacted pathways: A comprehensive review and assessment, Genome Biol. 20 (2019), 203.
[36] E.C. de Bruin, N. McGranahan, R. Mitter, et al., Spatial and temporal diversity in genomic instability processes defines lung cancer evolution, Science 346 (2014) 251–256.

[37] Y. Chen, W. Tang, H. Lin, et al., Wait-and-see treatment strategy could be considered for lung adenocarcinoma with special pleural dissemination lesions, and low genomic instability correlates with better survival, Ann. Surg. Oncol. 27 (2020) 3808–3818.
[38] P. Maione, A. Sgambato, F. Casaluce, et al., The role of the antiangiogenetic ramucirumab in the treatment of advanced non small cell lung cancer, Curr. Med. Chem. 24 (2017) 3–13.
[39] S. Negrini, V.G. Gorgoulis, T.D. Halazonetis, Genomic instability — An evolving hallmark of cancer, Nat. Rev. Mol. Cell Biol. 11 (2010) 220–228.
[40] L.R. Ferguson, H. Chen, A.R. Collins, et al., Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition, Semin. Cancer Biol. 35 (2015) S5–S24.
[41] R. Anandakrishnan, R.T. Varghese, N.A. Kinney, et al., Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations, PLoS Comput. Biol. 15 (2019), e1006881.
[42] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, et al., Signatures of mutational processes in human cancer, Nature 500 (2013) 415–421.
[43] L.B. Alexandrov, Y.S. Ju, K. Haase, et al., Mutational signatures associated with tobacco smoking in human cancer, Science 354 (2016) 618–622.
[44] S. Weissman, K. Takakura, G. Eibl, et al., The diverse involvement of cigarette smoking in pancreatic cancer development and prognosis, Pancreas 49 (2020) 612–620.
[45] S. Sidharthan, S. Kottilil, Mechanisms of alcohol-induced hepatocellular carcinoma, Hepatol. Int. 8 (2014) 452–457.
[46] M.S. Carlino, J. Larkin, G.V. Long, Immune checkpoint inhibitors in melanoma, Lancet 398 (2021) 1002–1014.
[47] P. Gotwals, S. Cameron, D. Cipolletta, et al., Prospects for combining targeted and conventional cancer therapy with immunotherapy, Nat. Rev. Cancer 17 (2017) 286–301.
[48] F. Battaglin, M. Naseem, H.J. Lenz, et al., Microsatellite instability in colorectal cancer: Overview of its clinical significance and novel perspectives, Clin. Adv. Hematol. Oncol. 16 (2018) 735–745.
[49] M. Ratti, A. Lampis, J.C. Hahne, et al., Microsatellite instability in gastric cancer: Molecular bases, clinical perspectives, and new treatment approaches, Cell. Mol. Life Sci. 75 (2018) 4151–4162.
[50] P.A.J. Muller, K.H. Vousden, p53 mutations in cancer, Nat. Cell Biol. 15 (2013) 2–8.
[51] R. Roskoski Jr., Targeting oncogenic Raf protein-serine/threonine kinases in human cancers, Pharmacol. Res. 135 (2018) 239–258.
[52] R. Nusse, H. Clevers, Wnt/β-catenin signaling, disease, and emerging therapeutic modalities, Cell 169 (2017) 985–999.
[53] D.K. Simanshu, D.V. Nissley, F. McCormick, RAS proteins and their regulators in human disease, Cell 170 (2017) 17–33.
[54] G. Hoxhaj, B.D. Manning, The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism, Nat. Rev. Cancer 20 (2020) 74–88.
[55] K. Ruan, G. Song, G. Ouyang, Role of hypoxia in the hallmarks of human cancer, J. Cell. Biochem. 107 (2009) 1053–1062.
[56] P. Zhao, H. Zhen, H. Zhao, et al., Identification of hub genes and potential molecular mechanisms related to radiotherapy sensitivity in rectal cancer based on multiple datasets, J. Transl. Med. 21 (2023), 176.
[57] E. Hessmann, G. Schneider, V. Ellenrieder, et al., MYC in pancreatic cancer: Novel mechanistic insights and their translation into therapeutic strategies, Oncogene 35 (2016) 1609–1618.
[58] M. Wirth, S. Mahboobi, O.H. Krämer, et al., Concepts to target MYC in pancreatic cancer, Mol. Cancer Ther. 15 (2016) 1792–1798.
[59] N. Li, L. Qiu, C. Zeng, et al., Bioinformatic analysis of differentially expressed genes and pathways in idiopathic pulmonary fibrosis, Ann. Transl. Med. 9 (2021), 1459.
[60] C. Wang, Q. Yu, T. Song, et al., The heterogeneous immune landscape between lung adenocarcinoma and squamous carcinoma revealed by single-cell RNA sequencing, Signal Transduct. Target. Ther. 7 (2022), 289.
[61] M. Yu, Y. Tian, M. Wu, et al., A comparison of mRNA and circRNA expression between squamous cell carcinoma and adenocarcinoma of the lungs, Genet. Mol. Biol. 43 (2020), e20200054.
[62] Y. Zhang, S. Li, Z. Lyu, et al., The co-mutation of EGFR and tumor-related genes leads to a worse prognosis and a higher level of tumor mutational burden in Chinese non-small cell lung cancer patients, J. Thorac. Dis. 14 (2022) 185–193.
[63] G. Lamberti, E. Andrini, M. Sisi, et al., Beyond EGFR, ALK and ROS1: Current evidence and future perspectives on newly targetable oncogenic drivers in lung adenocarcinoma, Crit. Rev. Oncol. Hematol. 156 (2020), 103119.
[64] V. Relli, M. Trerotola, E. Guerra, et al., Abandoning the notion of non-small cell lung cancer, Trends Mol. Med. 25 (2019) 585–594.
[65] V. Relli, M. Trerotola, E. Guerra, et al., Distinct lung cancer subtypes associate to distinct drivers of tumor progression, Oncotarget 9 (2018) 35528–35540.