BMC Cancer

RESEARCH

Open Access

Lung Cancer Biomarker Database (LCBD): a comprehensive and curated repository of lung cancer biomarkers

Yinghong Li[1*†], Zhuohao Tong[1†], Yinqi Yang[1], Yu Wang[1], Lu Wen[1], Yuke Li[1], Mingze Bai[1], Yongfang Xie[1], Bo Li[2*] and Kunxian Shu[1*]

Abstract

Background Lung cancer remains a leading cause of cancer-related mortality, primarily because of the lack of effective diagnostic and therapeutic biomarkers. To address the issue of fragmented biomarker data across numerous publications, we have developed the Lung Cancer Biomarker Database (LCBD, http://lcbd.biomarkerdb.com).

Methods We comprehensively reviewed biomarker-related studies up to June 30, 2023, and extracted relevant biomarker information. The identified biomarkers were systematically annotated, including genes, proteins, GO terms, KEGG pathways, biomarker types, molecular types, developmental stages, discovery methods, sources, populations, and sample sizes. The LCBD online platform was developed to integrate lung cancer biomarker data, and provide search, browsing, and data download functions for researchers. To validate the data in the LCBD, we conducted three case studies comparing models with and without LCBD data.

Results After deduplication and summarization, we collected 1,447 unique biomarkers that were systematically annotated. We then developed the LCBD specifically for use in lung cancer diagnosis. The validity of the biomarkers in the LCBD was confirmed using prognostic models, diagnostic models, and immune infiltration models.

Conclusion The LCBD provides a centralized platform for lung cancer biomarkers, facilitating early screening and personalized treatment. This database is poised to become a valuable resource for lung cancer research and therapeutic strategies.

Keywords Lung cancer, Biomarker database, Prognostic model, Immune infiltration model

Introduction

Lung cancer remains a major global health challenge because of its high incidence and mortality rates [1]. According to a report by Rebecca L. Siegel [2], it is anticipated that by 2024, there will be 234,580 new cases of lung cancer in the United States, accounting for 11.72% of all new cancer cases. Lung cancer is the leading cause of cancer-related death among individuals aged 50 years and above, resulting in more deaths than breast, prostate, and colorectal cancers combined [3]. Often, lung cancer does not manifest with specific symptoms in its early stages, as a result, most patients

†Yinghong Li and Zhuohao Tong contributed equally to this work.

*Correspondence:
Yinghong Li
liyinghong@cqupt.edu.cn
Bo Li
libcell@cqnu.edu.cn
Kunxian Shu
shukx@cqupt.edu.cn
[1] Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China
[2] College of Life Sciences, Chongqing Normal University, Chongqing 401331, P. R. China

are diagnosed at advanced stages or with distant metastases [4], missing the critical period for effective treatment. Although there has been a decrease in the lung cancer death rate in the United States in recent years, the five-year survival rate remains below 25% [3], likely due to these late diagnoses. Early detection and treatment of lung cancer could significantly increase the five-year survival rate to approximately 60% [5].

The current battery of lung cancer diagnostics encompasses a range of techniques such as chest X-ray, computed tomography (CT), positron emission tomography (PET), biopsy, and biomarker detection. Although commonly used, chest X-rays have limited efficacy in identifying small or deeply situated lung tumours; CT scans offer greater sensitivity at the cost of increased radiation exposure [6]. Although adept at detecting minute tumours, PET scans have the disadvantages of even greater radiation doses and significant expense [7]. Biopsies remain the gold standard for confirming a cancer diagnosis, but their invasive nature renders them impractical for use in early detection and widespread screening [8]. Biomarker detection offers a quicker, less invasive alternative, but the current clinical biomarkers, such as CEA and NSE have low sensitivity and specificity and are affected by various confounding factors [9]. These deficiencies have spurred research into biomarkers with increased sensitivity and specificity.

In recent years, with the development of precision medicine, scholars have focused on lung cancer biomarkers, hoping to identify the risk factors for lung cancer, predict effect of the treatment, and improve the diagnosis and treatment outcomes through these biomarkers. Currently, a range of lung cancer biomarkers are utilized in clinical practice and research, some of which, including epidermal growth factor receptor (EGFR), human epidermal growth factor receptor 2 (HER2), and KRAS (KRAS proto-oncogene, GTPase), have significant importance and value [10]. EGFR is a gene that is overexpressed or mutated in lung cancer cells and is closely associated with the development, progression, and drug resistance of the disease. Similarly, the mutation of HER2 in lung cancer cells is correlated with tumour aggressiveness and prognosis. In addition, KRAS is involved in lung cancer cell proliferation and metastasis. These genes can serve as valuable biomarkers for guiding targeted drug therapy selection for patients with lung cancer on the basis of different genotypes and function as predictive biomarkers to assess treatment response and prognosis for patients. These studies highlight the importance of comprehensive biomarker analysis in facilitating diverse and personalized approaches to cancer diagnosis and treatment while advancing progress in these areas.

The academic community has extensively documented an array of lung cancer biomarkers, detailing their diverse types and functions across numerous publications. Lung cancer research currently struggles with biomarker data dispersed across diverse sources lacking a centralized repository. This hinders their efficient discovery and application. Therefore, the creation of a meticulously curated and comprehensive database for lung cancer biomarkers, that offers streamlined access to refined biomedical information, is both urgent and essential.

Although there are several resources containing lung cancer biomarkers, such as the Lung Cancer Metabolome Database (LCMD) [11], the Lung Cancer Circular RNA Biomarker Database (LCcircDB) [12], The Marker [13], MethMarkerDB [14], and MarkerDB [15], these resources fall short in terms of the diversity and quantity of biomarkers. Specifically, LCcircDB is limited to circRNA biomarkers, overlooking vital information on genes, proteins, and metabolites. The LCMD, which draws from only 65 mass spectrometry-based lung cancer metabolomics papers, offers a limited selection of metabolite biomarkers that is insufficient for reliable lung cancer diagnostics and prognosis. MethMarkerDB includes only 379 metabolites and fails to differentiate between lung cancer subtypes. Both the LCMD and LCcircDB are based on pre-2013 literature and lack the recent updates necessary to reflect the latest findings in lung cancer biomarker research. The MarkerDB's coverage of lung cancer biomarkers are inadequate, listing only 96 types. Despite having a significant number of biomarkers, TheMarker relies heavily on the results of differential expression analysis from a few transcriptomic datasets (GSE55859, GSE126044, TCGA-LUAD, and TCGA-LUSC) and uses only the criteria of upregulation and downregulation, which is an imprecise method for defining biomarkers. Additionally, these databases offer limited biomedical annotations, often lacking detailed analyses and extensive annotation of biomarker functions, pathways, and genetic locations.

Therefore, the creation of an updated and detailed lung cancer biomarker database integrating a broad spectrum of biomarker types and functionalities alongside extensive biomedical data is imperative. The purpose of the envisioned LCBD is to enhance our understanding and application of these biomarkers and to bolster the development of innovative therapeutic and diagnostic approaches.

Materials and methods

The development of the LCBD involved three principal stages: the collection of data (including biomarkers and biomarker panels), the annotation of these biomarkers, and the development of the database's website.

Additionally, to substantiate the validity and assess the robustness of the data, three case studies were performed to build prognostic prediction models, diagnostic models, and immune infiltration models using the completed LCBD, as shown in Fig. 1.

Identification and curation of lung cancer-related biomarkers from literature published prior to June 2023

To systematically gather data on lung cancer biomarkers, comprehensive literature searches were conducted using specific keywords. For PubMed, the keywords ("lung cancer" OR "lung carcinoma" OR "pulmonary carcinoma") AND "biomarker" were searched in titles or abstracts. Google Scholar searches were performed with combinations such as "lung cancer + biomarker," "lung carcinoma + biomarker," and "pulmonary carcinoma + biomarker," and all searches were limited to studies published before June 30, 2023. The following details were extracted from the identified articles: biomarker denomination, genes associated with the biomarker, role of the biomarker (including but not limited to diagnosis, prediction, prognosis, and pharmacodynamics), molecular classification (e.g., miRNA, gene, protein, lncRNA, circRNA and metabolite), validity, methods of validation, conditions of application, source material (such as tissue, blood, plasma or serum), stage of research (whether investigational, approved, or clinical), cancer subtype (including non-small cell lung cancer, adenocarcinoma, squamous cell lung cancer and large cell lung cancer, etc.), demographic data, sample sizes, and the association of pharmacodynamic biomarkers with specific therapeutic agents when available. Additionally, we have systematically recorded the clinical trial IDs. For each biomarker, our selection criteria were as follows: a biomarker was included in the LCBD only if the study authors explicitly designated a specific molecule and provided conclusive evidence of its association with lung cancer in their research findings. Furthermore, biomarkers that were experimentally validated through multiple independent studies and demonstrated clinical utility in disease diagnosis, prognosis evaluation, or therapeutic response prediction with cross-referenced documentation of supporting evidence from diverse studies were prioritized for inclusion.

Comprehensive annotation of biomarkers

To facilitate a better understanding and practical application of lung cancer biomarker information, we meticulously annotated and analysed the biomarkers within the LCBD from various perspectives, including analysing biomarkers from multiple dimensions and employing various methodologies tailored to their molecular characteristics. (1) Gene biomarkers were annotated for function and pathway associations utilizing the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases, respectively. (2) For protein

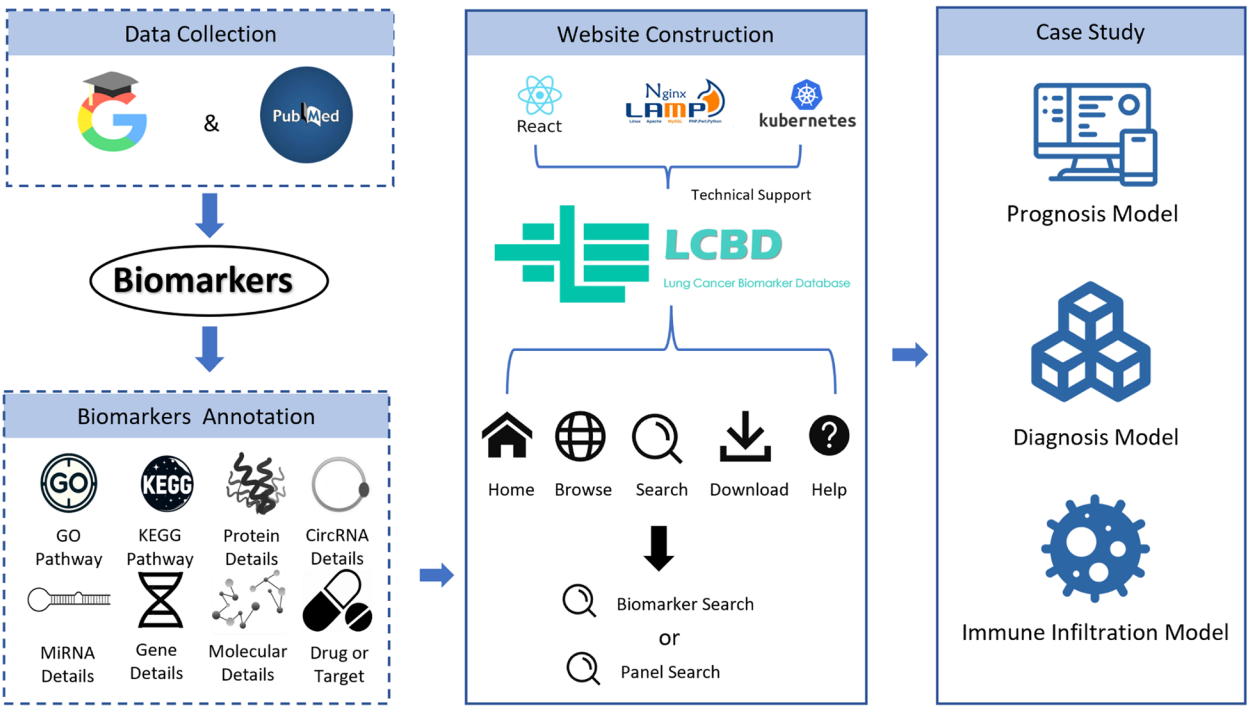


Fig. 1 Flowchart illustrating the development process of the LCBD

biomarkers, we extracted detailed annotations, including protein names, functions, interaction networks, and other critical information, from the UniProt database [16]. (3) Annotations for microRNAs, lncRNAs, and circRNAs were sourced from specialized repositories—mirBase [17], LNCipedia [18] and Circbank [19], respectively—which provide molecular nomenclature and sequence data. For intermolecular interactions, we explored databases such as LncCeRBase [20], Mirtarbase [21], and NPInter [22] and compiled evidence of interactions between microRNAs and genes, between microRNAs and lncRNAs, and between lncRNAs and proteins. Predictive biomarkers were specifically annotated with drug-related information, including drug names, targets, and clinical indications—from resources such as the DrugBank [23] and the Therapeutic Target Database (TTD) [24]. Moreover, we included annotations for biomarker types, development stages, biomarker discovery methods, sample sources, populations, sample sizes, and other pertinent details regarding biomarkers.

Development of the LCBD

The architecture of the LCBD was designed with a browser/server (B/S) structure, with clear separation between its frontend and backend, utilizing contemporary technologies for an efficient and robust user experience. For the frontend, we selected React from among the popular trio of Angular, Vue, and React, as a framework to create a responsive design that adapts seamlessly to various screen sizes and resolutions, thus facilitating access on a wide range of devices from mobile phones to desktop computers. On the backend, LCBD employed the LNMP stack [25], a robust framework of Linux, Nginx, MySQL, and PHP, augmented by Redis caching and ElasticSearch, to streamline the biomarker and panel search process. The search functionality was optimized using ElasticSearch [26], leveraging techniques such as word segmentation and inverted indexing to enhance accuracy. These methods enabled precise and efficient retrieval, even when processing large-scale datasets and complex queries.

To ensure high system availability, scalability, and fault tolerance, Kubernetes [27] was chosen as the deployment platform. Kubernetes provides comprehensive container orchestration features, including automated scaling, load balancing, self-healing, and rolling updates. These capabilities allowed the LCBD to maintain continuous availability, dynamically allocate resources in response to demand, and recover seamlessly from system failures. The selection of Kubernetes over other container orchestration platforms, such as Docker Swarm or Apache Mesos, was based on its advanced feature set, robust ecosystem, and widespread industry adoption, ensuring a reliable and future-proof deployment environment.

Datasets for case studies

To ascertain the validity of the LCBD, we conducted three case studies focusing on its predictive capabilities. Data for the database prediction models were sourced from Xena (https://xenabrowser.net/datapages/) [28], from which we downloaded RNA-seq datasets for lung adenocarcinoma from the TCGA database. Both the diagnostic model and the immune cell infiltration prediction model utilized clinical and gene expression data on lung adenocarcinoma (LUAD) from the TCGA (https://cancergenome.nih.gov, accessed on August 1, 2022). The final datasets comprised 1,214 independent samples, each with a complete set of clinical information and gene expression profiles. In addition, proteomic data from the CPTAC database, which includes 106 independent samples, were obtained for further refinement of the lung cancer diagnostic model.

Identification of differentially expressed genes (DEGs) and differentially expressed proteins (DEPs)

To mitigate variations arising from sample preparation, storage, or sequencing processes, we normalized the expression profile data using the "edgeR" package within R software, version 4.0.3. The "limma" package was then employed to identify DEGs and DEPs by comparing normal lung samples against LUAD samples. The selection criteria for DEGs were a false discovery rate (FDR) of less than 0.05 and an absolute log2 fold change greater than 2. We utilized the "p.adjust" function to calculate the FDR for each gene, thereby establishing significant q values, with an FDR q value of less than 0.05 deemed statistically significant. To visualize differential gene expression, we generated volcano plots and heatmaps using the "ggplot2" and "pheatmap" packages, respectively.

Analysis of case studies

On the basis of studies conducted by Zhou Jing [29], Suli Liu [30], and Lingge Yang [31], we developed three distinct predictive models tailored to assess the prognosis, diagnosis, and classification of immune cell infiltration types in lung cancer patients. Through these three case studies, we not only explore the practical applications of LCBD's lung cancer biomarkers but also compare them with other publicly accessible lung cancer databases, providing an in-depth analysis of LCBD's advantages. Model performance was evaluated using the K-fold cross-validation method with $K=5$.

Development of a prognostic model for lung cancer

To identify survival-related DEGs, we performed univariate Cox proportional hazards analyses, at a the significance level of $P < 0.01$ [29]. Risk scores for each sample were calculated within the prognostic model according to the expression levels of these DEGs and their corresponding coefficients, as defined by the following formula:

$$\text{RiskScore} = \text{coef}_1 \times \text{expression}_A + \text{coef}_2 \times \text{expression}_B + \cdots + \text{coef}_N \times \text{expression}_N$$

where $\text{coef}_{1\text{ to }N}$ denotes the survival-related coefficients for the DEGs, and $\text{expression}_{A\text{ to }N}$ represents the expression levels of the DEGs. The patients from the TCGA lung cancer dataset were stratified into high-risk and low-risk groups according to the median risk score. Differences in survival between these groups were analyzed, via Kaplan-Meier survival analysis with the "survival" and "survminer" packages in R.

Furthermore, we investigated the relationships between clinicopathological features and risk scores via univariate Cox regression analysis, via the Survival package in R. The accuracy of the survival outcome predictions based on clinicopathological factors and risk scores was evaluated by constructing time-dependent receiver operating characteristic (ROC) curves using the "survivalROC" R package.

Development of a diagnostic model for lung cancer

The Kullback–Leibler (KL) divergence [32], is a nonsymmetric value that quantifies the divergence between two probability distributions for a variable i, as defined by the following formula:

$$D_{kL} = -\sum_{i=1}^{n} P(i) * \ln\left(\frac{Q(i)}{P(i)}\right)$$

where P denotes the actual data distribution, and Q represents a theoretical or estimated approximation of P.

Gene expression data distributions are readily ascertainable from a small dataset. The disparity between any two gene expression distributions can be quantified using the KL divergence. Similar distributions suggest a lack of association between the genes. The use of the KL divergence has been shown to be effective in identifying essential DEGs for diagnosis [33].

We devised a diagnostic model using a deep neural network endowed with two hidden layers (128 and 64 neurons), selecting focal loss as the loss function. Model performance was assessed using the K-fold cross-validation method with a K of 5. The dataset was split into 80% training and 20% test sets. Within the training set, we implemented fivefold cross-validation, randomly dividing it into five subsets to iteratively assess model performance. Additionally, the batch size was set to 32, meaning that 32 samples are processed per model update, a choice informed by computational efficiency and common practice in similar studies. Additionally, we employed the Adam optimizer, with its learning rate set to 0.0005, to balance stability and convergence speed during training. To assess the model's predictive accuracy, we generated an ROC curve on the test sets. Furthermore, performance metrics such as accuracy, recall, and precision were employed to evaluate the efficacy of the models with and without gene selection. These metrics were computed from the confusion matrix.

Development of an immune infiltration model

Using the "GSVA" software package (v1.42.0) [34] and single-sample gene set enrichment analysis (ssGSEA) data, we calculated immune scores for 28 distinct immune cell types, grouped by their functions in antitumour immunity, protumour activities, and immunosuppression, among others. Following normalization of the immune cell enrichment score matrix, we conducted NMF classification to determine the extent of immune infiltration in patients. The sequence was 2:6, the method was Brunet, and the nrun value was 30.

Next, we applied k-nearest neighbour (KNN)(n_neighbors = 50) [35] and random forest(n_estimators = 100) [36] classifications to predict immune cell infiltration types. The dataset was divided into 70% training and 30% test sets. Within the training set, we performed fivefold cross-validation to train and optimize KNN and random forest classifiers. The final model's performance was assessed on the test set using the area under the ROC curve (AUC).

Inclusion of comparative databases in the case studies

To date, publicly available lung cancer biomarker databases include LCMD [11], LCcircDB [12], TheMarker [13], MethMarkerDB [14, 15]. Because LCMD provides only metabolomics data and does not include transcriptomics-based biomarkers, it was excluded from our comparative analyses and the subsequent case studies. In addition, LCcircDB was excluded due to its current inaccessibility.

Thus, for our case studies, we systematically compared LCBD, TheMarker, MethMarkerDB, and MarkerDB. It should be noted that most of the data in TheMarker

are derived from differentially expressed genes identified through transcriptome datasets (e.g., GSE55859, TCGA-LUAD), and there is ongoing debate regarding the definition of lung cancer biomarkers using this approach. Therefore, we selected only those biomarkers from TheMarker that were supported by clinical or experimental evidence to construct our model. Furthermore, from MethMarkerDB and MarkerDB, we chose lung cancer–related biomarkers for building our case study models (prognostic, predictive, and immune infiltration models).

Results and discussion

Quantification of biomarkers

An extensive literature review was conducted to generate a comprehensive dataset comprising 3,175 biomarkers for this study. This dataset encompasses biomarkers with a diverse range of applications, including 176 biomarkers for detection, 1,409 for diagnosis, 600 for prognosis, and 591 for prediction (Table 1). In terms of molecular categorization, the dataset included 601 microRNAs, 979 proteins, 88 circular RNAs (circRNAs), and 742 genes (Table 2). Among these biomarkers, 743 have been approved, 236 are in the clinical stage, and 924 are in the research stage (Table 3).

Exploring the interface and features of the LCBD

The LCBD provides users with two distinct search functionalities: "quick search" and "advanced search". The "quick search" option, which is accessible directly from the homepage, enables users to efficiently locate specific biomarkers of interest (Fig. 2a). In contrast, the "advanced search" feature allows for more refined queries (Fig. 2b). The LCBD facilitates biomarker searches on the basis of a comprehensive array of criteria, including but not limited to designation, type, molecular characteristics, associated medical conditions, application methods, source, developmental stage, disease subtypes, demographic details of the population, and size of the study sample. Following the execution of either search mode, the database generates an organized table. This table lists the identified biomarkers and is designed for efficient browsing and sorting. Embedded within this table are hyperlinks, each leading to detailed information pertinent to the respective search terms.

Selecting a biomarker name (for example, HER2) directs the user to a detailed page about the biomarker, divided into five distinct sections: (I) Basic information about the biomarker, including biomarker ID, name, gene ID, and corresponding UniProt database ID (Fig. 3a); (II) detailed information from a range of literature sources, such as molecular types, statuses, identification methods, origins, developmental stages as biomarkers, and applicable lung cancer subtypes (Fig. 3b); (III) biomarker-related drugs and their targets, including a table sourced from the DrugBank [23] and TTD databases [24] (Fig. 3c); (IV) information on biomarkers classified by molecular type, such as genes, proteins, lncRNAs, miRNAs, and circRNAs, compiled from sources such as miRBase, LNCipedia, and Circbank (Fig. 3d); and (V) annotation information for biomarkers, including GO terms and KEGG pathways (Fig. 3e).

The "Panel Search" module allows users to search for biomarker panels according to their name and type, producing a list that is both browsable and sortable and contains comprehensive information related to the search keyword (Fig. 4a). Clicking on "Panel ID" redirects users to a page with detailed information about the selected lung cancer biomarker, organized into three sections (Fig. 4b): (I) Basic information about the panel, (II) detailed information about the biomarkers included in the panel, and (III) a reference section.

Additionally, users can explore the LCBD in depth via the “Browse” function. Within the LCBD, biomarkers can be filtered using five criteria—biomarker ID, biomarker name, biomarker type, drug name, and disease subtype (Fig. 5a). Clicking a highlighted biomarker ID directs users to a page containing detailed information about the selected biomarker (Fig. 3). Similarly, on the “Panel Browse” page, biomarkers can be filtered according to three criteria—biomarker Panel ID, biomarker Panel name, and biomarker Panel type (Fig. 5b). Clicking a highlighted biomarker ID on this page likewise directs users to a page with detailed information about the selected biomarker (Fig. 4).

Users can upload their biomarker data to the LCBD (Supplementary Fig. 1) via a dedicated submission interface, where they may enter information on

Table 1 Quantification of different types of biomarkers

Type	Detective	Diagnostic	Prognostic	Predictive	Others
Number	176	1409	600	591	399

Table 2 Quantification of biomarkers of different molecular types

Type	MicroRNA	Protein	CircRNA	DNA	Gene
Number	601	979	88	158	742

Table 3 Quantification of biomarkers in different research stages

Stage	Approved	Clinical	Research
Number	743	236	924

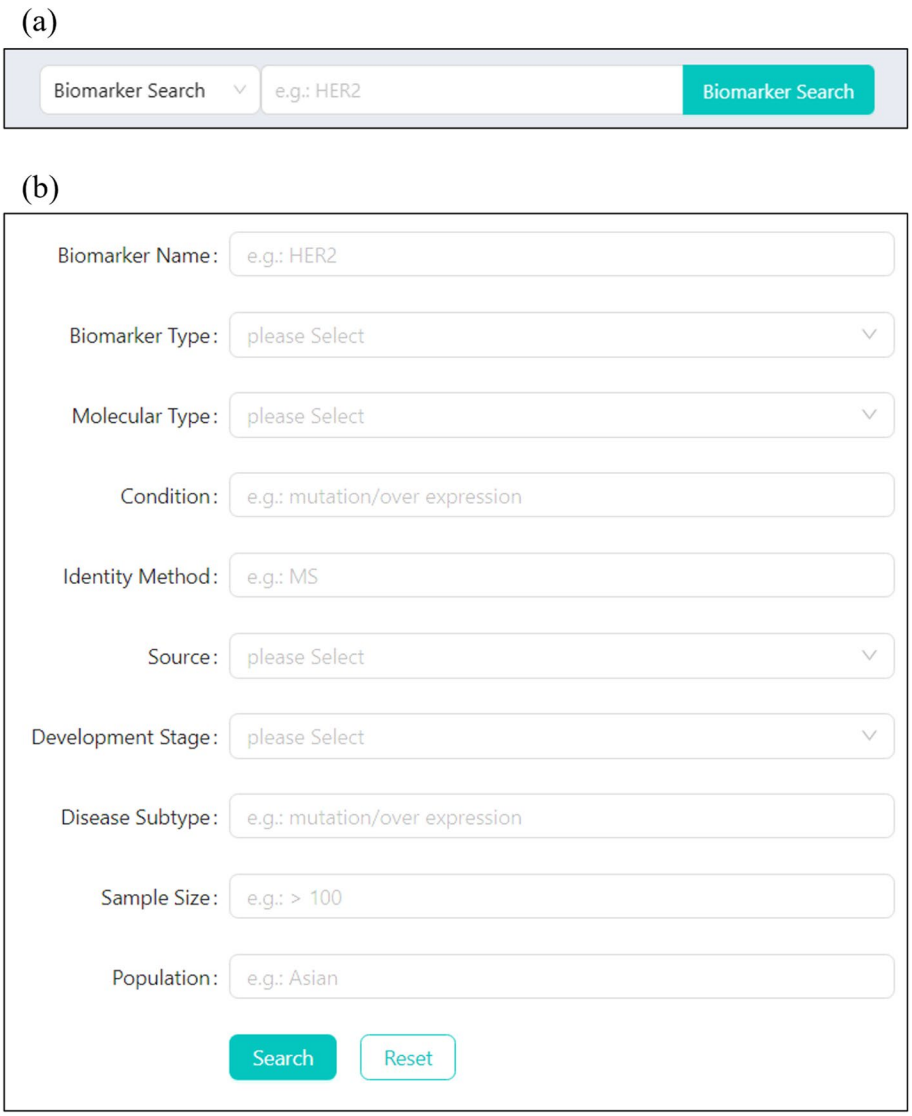


Fig. 2 Search module. **a** Quick search. This feature enables users to conduct searches by biomarker name, facilitating the exploration of research within the LCBD. **b** Advanced search. This sophisticated search tool allows users to navigate through LCBD research utilizing a set of ten filters

published or clinically validated lung cancer biomarkers according to their research needs. To ensure data accuracy and reliability, all submissions undergo a rigorous review and validation process; only upon approval are the corresponding data formally updated in the database and made publicly available.

Comparison with existing biomarker databases

Comprehensive databases featuring extensively annotated lung cancer biomarkers are fundamental to advancing lung cancer research. However, databases such as LCMD [11], LCcircDB [12], TheMarker [13], MethMarkerDB [14], and MarkerDB [15] currently lack adequate biomarker data and annotations, which impedes critical research in the diagnosis and treatment of lung cancer (Table 4).

Specialized databases such as LCMD and LCcircDB, which focus on particular types of lung cancer biomarkers (metabolites and circRNA), consistently ignore other necessary data. The limited data volume within these databases is insufficient to meet the scholarly demand for extensive biomarker datasets. Moreover, these databases fail to prioritize the collection of essential information concerning biomarker-related drugs, targets and biomarker panels, and their annotations lack the necessary variety and detail.

Recent biomarker databases such as TheMarker, MethMarkerDB, and MarkerDB, although broader in scope

(a)

Basic Info

Biomarker ID	LCBD-1099	Biomarker Name	HER2
Gene_ID	2064	Uniprot_ID	P04626

(b)

Biomarker information from different references

Biomarker Type	Molecular Type	Condition	Development Stage	Identity Method	Method Type	Source	Disease Subtype
Predictive	Gene	Mutation 2%	Approved	IHC (immunohistochemistry) and/ or FISH	Experimental Method	-	Non-Small-Cell Lung
Predictive	Gene	HER2 gene copy number increase	Research	-	-	-	Non-Small-Cell Lung
Diagnostic	Gene	Sensitivity = 45%– 64% Specificity = 85%	Research	ELISA	Experimental Method	-	Lung Cancer

(c)

Drug Name	DrugBank Accession Number	DrugBank Drug Type	DrugBank Drug Synonyms	DrugBank Drug External IDs
Afatinib [1]	DB08916	Small Molecule	-	BIBW 2992, BIBW-2992, BIBW2992

Target Name	DrugBank ID	Target Gene ID	TTD Target ID	DrugBank Target ID
Epidermal growth factor receptor	-	1956	T59328	P00533
Receptor tyrosine-protein kinase erbB-2	-	2064		P04626
Receptor tyrosine-protein kinase erbB-4	-	2066		Q15303
Erbb2 tyrosine kinase receptor (HER2)	-	2064	T14597	

(d)

Gene Ontology

GO ID	GO Terms	GO Category
GO:0005886	plasma membrane	cellular component
GO:0016021	integral component of membrane	cellular component
GO:0004713	protein tyrosine kinase activity	molecular function
GO:0008284	positive regulation of cell population proliferation	biological process

KEGG Pathways

Pathway Database	Pathway ID	Pathway Name
KEGG	hsa05200	Pathways in cancer - Homo sapiens (human)
KEGG	hsa05205	Proteoglycans in cancer - Homo sapiens (human)
KEGG	hsa05212	Pancreatic cancer - Homo sapiens (human)
KEGG	hsa05213	Endometrial cancer - Homo sapiens (human)
KEGG	hsa05223	Non-small cell lung cancer - Homo sapiens (human)

(e)

References List

[1] Shoshan-Barmatz V, Bishitz Y, Paul A, et al. A molecular signature of lung cancer: potential biomarkers for adenocarcinoma and squamous cell carcinoma. Oncotarget. 2017;8(62):105492-105509. Published 2017 Nov 6. doi:10.18632/oncotarget.22298

[2] Melosky B, Blais N, Cheema P, et al. Standardizing biomarker testing for Canadian patients with advanced lung cancer. Curr Oncol. 2018;25(1):73-82. doi:10.3747/co.25.3867

[3] Jain D, Nambirajan A, Borczuk A, et al. Immunocytochemistry for predictive biomarker testing in lung cancer cytology. Cancer Cytopathol. 2019;127(5):325-339. doi:10.1002/cncy.22137

[4] Wishart DS, Bartok B, Oler E, et al. MarkerDB: an online database of molecular biomarkers. Nucleic Acids Res. 2021;49(D1):D1259-D1267. doi:10.1093/nar/gkaa1067

[5] Huang Y, Cho HJ, Stranger BE, Huang RS. Sex dimorphism in response to targeted therapy and immunotherapy in non-small cell lung cancer patients: a narrative review. Transl Lung Cancer Res. 2022;11(5):920-934. doi:10.21037/tlcr-21-1013

[6] Megan Baumgart & Kishan Pandya?(2016)?The use of biomarkers in the treatment of non-small cell lung cancer,?Expert Review of Precision Medicine and Drug Development,?1:1,?25-36,?DOI:?10.1080/23808993.2016.1136558

Fig. 3 Biomarker details page. **a** Basic information, including biomarker ID, name, and gene ID. **b** Biomarker data from diverse literature sources. **c** Associated drugs and target details. **d** Information categorized by molecular type: genes, proteins, lncRNAs, miRNAs, and circRNAs. **e** Gene Ontology and KEGG pathway annotations

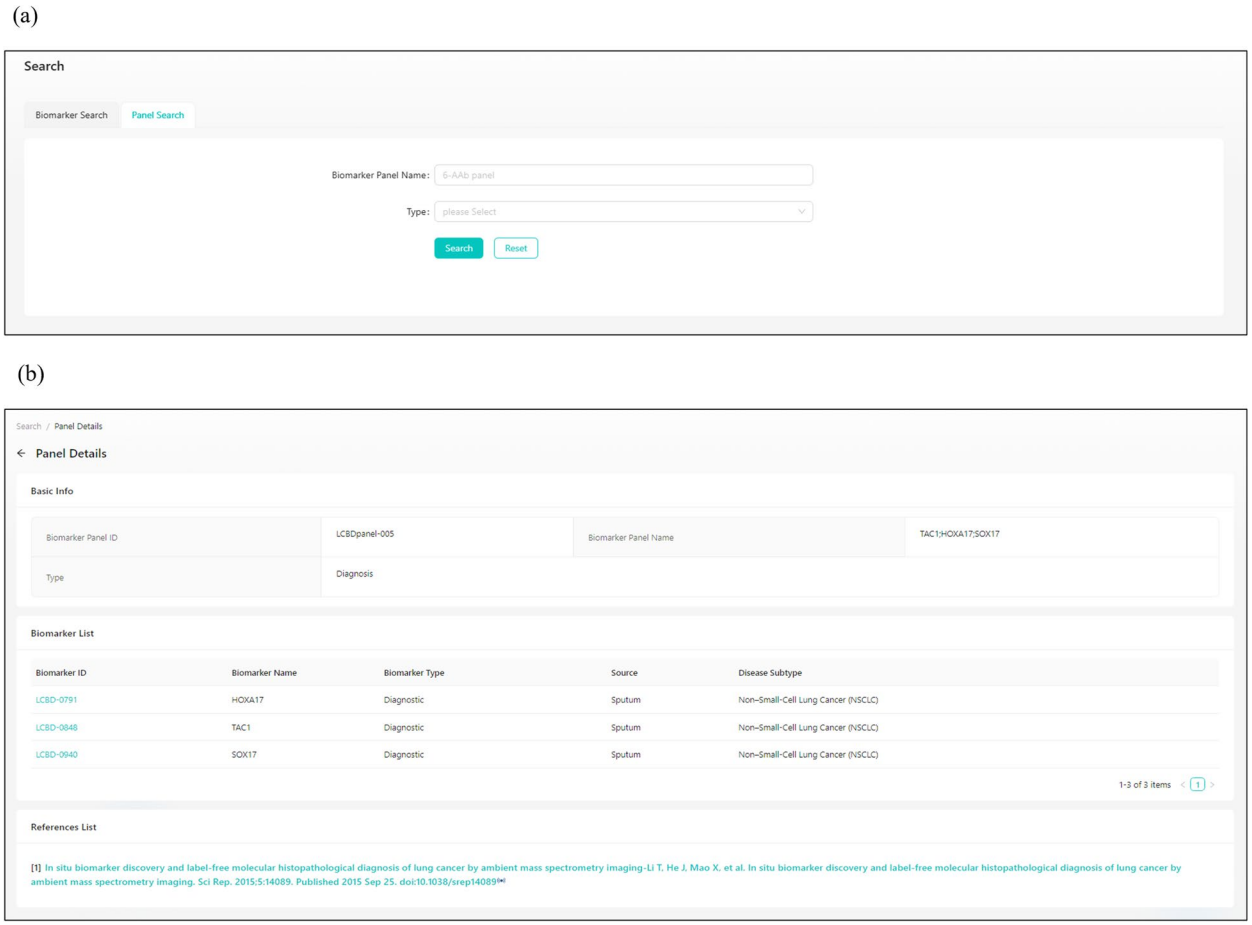
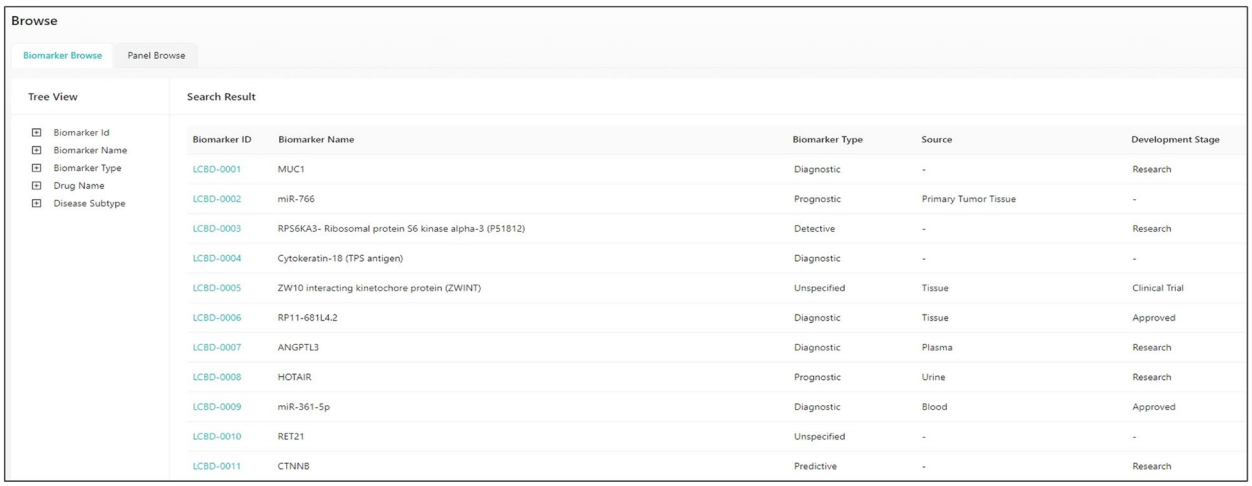


Fig. 4 Biomarker panel search and detail overview. **a** Panel search module: Users can search for study panels using filters for name and type. **b** Panel details page (here chose TAC1, HOXA17, and SOX17 as an example)

than the LCBD, do not match our database in capturing the extensive range and quality of lung cancer biomarkers. MarkerDB, for example, catalogues only certain lung cancer subtypes, such as somatic adenocarcinoma of the lung and small cell lung cancer, omitting other crucial NSCLC subtypes, such as squamous cell cancer and large cell cancer, and contains only 96 biomarkers sourced from 136 publications. MethMarkerDB includes only 379 biomarkers from 218 publications without detailed subtype annotations. Although TheMarker includes a significant number of biomarkers, these biomarkers are derived primarily from differential expression analyses of a limited number of transcriptome datasets (GSE55859, GSE126044, TCGA-LUAD, and TCGA-LUSC), and the validity of this method in defining lung cancer biomarkers has been debated. Moreover, TheMarker emphasizes mRNAs and lncRNAs; neglects miRNA and protein data, including only 9 miRNAs and 46 proteins; and generally labels biomarkers under the broad category of "Lung Cancer" without assigning them to specific lung cancer subtypes.

To further demonstrate the superiority of LCBD data, we conducted a statistical analysis of LCBD alongside other accessible databases, focusing on each database's data inclusion capacity, coverage, and statistical significance. Specifically, based on five high-quality lung cancer biomarker studies published between February 2024 and August 2024 in authoritative journals, we systematically integrated 69 lung cancer biomarker datasets validated through multi-center clinical trials (Supplementary Table 1). We then constructed a contingency table for database inclusion status (included/not included) across five major biomarker databases (LCBD, MarkerDB, TheMarker, MethMarkerDB, and LCMD) and performed a chi-square test of independence. The results showed $\chi^2(4) = 127.51$ ($p = 1.33 \times 10^{-26}$, adjusted using the Benjamini–Hochberg method; Cramer's $V = 0.610$), confirming statistically significant heterogeneity in biomarker inclusion among the databases ($\alpha = 0.01$).

(a)



(b)

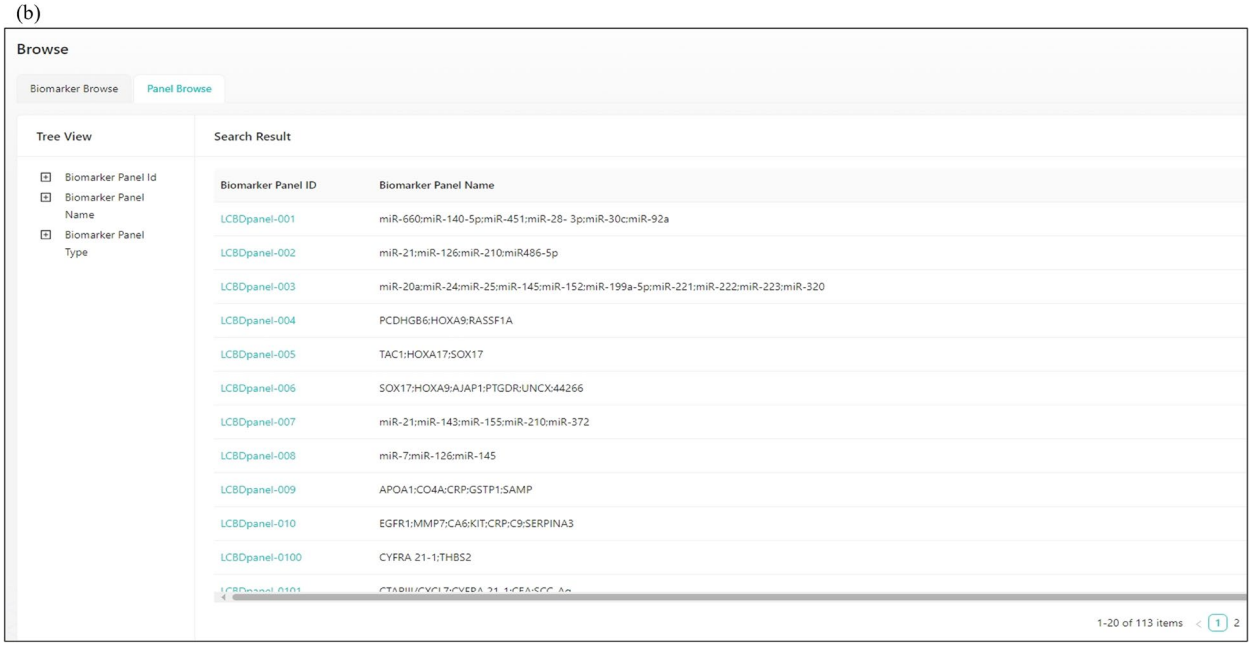


Fig. 5 Browse module. **a** Users can filter biomarkers in the LCBD using five criteria: Biomarker ID, Biomarker Name, Biomarker Type, Drug Name, and Disease Subtype. **b** Users can filter biomarker panels in the LCBD using three criteria: Biomarker Panel ID, Biomarker Panel Name, and Biomarker Panel Type

Further analysis revealed that LCBD achieved a coverage rate of 78.26%, which is markedly higher than the mean coverage of the other databases (16.23% ± 13.72%) ($t = -9.03$, $p < 0.01$, Cohen's $d = 4.52$). Moreover, the adjusted odds ratio (OR = 18.32, 95% CI: 9.51–35.28) indicates a strong association between LCBD and the 69 biomarkers, and the F1 score reached 0.64 (Table 5). Under a controlled false discovery rate (FDR < 5%), these findings establish LCBD as a preferred database that combines high coverage (breadth) with high precision (quality), serving as a benchmark data platform for translational research on lung cancer biomarkers.

Conversely, the LCBD contains an impressive collection of 1447 lung cancer biomarkers, covering a wide spectrum of biomarker types (prognosis, detection, diagnosis, monitoring, treatment, therapy, and prediction) and molecular categories (DNA, gene, mRNA, protein, miRNA, lncRNA, circRNA, and chemical), thus providing lung cancer researchers with a comprehensive repository of biomarker data. Moreover, the LCBD provides

Table 4 Comparison of LCBD with other lung cancer biomarker databases

	LCBD	LCMD	LCcircDB	TheMarker	MethMarkerDB	MarkerDB
No. of Biomarkers	1447	2013	1029	2439	379	96
Major Molecular type	Non-single class[a]	Metabolome	CircRNA	Non-single class[b]	DNA methylation	Non-single class[a]
Major cancer subtypes	SCC; ADC; SCLC; LUSC; LCC; NSCLC; SQLC; LC	NSCLC; SCLC; LC	ADC; NSCLC; LC	LUSC; LUAD	LC	ADC; SCLC; LC
Major Biomarker Type	Prognosis; Detection; Diagnosis; Monitoring; Treatment; Therapy; Prediction	Diagnosis; Therapy	not yet	Pharamacodynamic; Safety; Monitoring; Predictive; Surrogate; Endpoint	Diagnosis; Prognosis	Diagnosis; Prognosis; Prediction; Exposure
Panel	included	included	not yet	not yet	not yet	included
Drug & target	included	not yet	not yet	included	not yet	not yet
Annotation	included	included	included	included	included	included
Sample Sizes	included	included	not yet	not yet	not yet	not yet

SCC squamous cell carcinoma, *ADC* adenocarcinoma, *SCLC* small cell lung cancer, *LUSC* lung squamous cell carcinoma, *LCC* large cell carcinoma, *NSCLC* non-small cell lung cancer, *SQLC* squamous cell lung cancer, *LC* lung cancer

[a] Multiple molecular types, including DNA, gene, mRNA, protein, miRNA, lncRNA, circRNA, and chemical

[b] Multiple molecular types, including DNA, mRNA, protein, miRNA, lncRNA, and chemical

supplementary information on biomarker-associated drugs and targets, enriching the biomarker dataset.

Case studies

Prognostic model for lung cancer

The effectiveness of the LCBD was corroborated by three distinct case studies, each utilizing different approaches to model construction. The first case study focused on the development of a prognostic model. Through the analysis of normal lung samples and lung adenocarcinoma samples, 745 DEGs were identified (Supplementary Fig. 2). A prognostic model for lung cancer patients was constructed through the use of a Cox regression model [37] to conduct separate Cox univariate analyses on both the control group (non-LCBD group, consisting of 745 DEGs) and the experimental group (LCBD group, consisting of 745 DEGs as well as the union of genes from the LCBD, totalling 1608 genes). Subsequently, lung cancer prognosis risk scores were calculated, and the top 10 genes significantly associated with prognosis ($P<0.05$) were selected for further risk score computation (Supplementary Table 2 and Supplementary Table 3).

Table 5 Statistical Comparison of LCBD with Other Databases

Database	Odds Ratio	F1 Score	Coverage Rate (%)
LCBD	18.32	0.642857	78.26087
LCMD	0.352314	0.119048	14.49275
TheMarker	0.149601	0.059524	7.246377
MethMarkerDB	1.535627	0.297619	36.23188
MarkerDB	0.149601	0.059524	7.246377

By utilizing the derived risk score, the optimal cut-off values for categorizing lung cancer patients into high-risk and low-risk groups (experimental group: 2.6; control group: 1.9) were determined. A prognostic model for lung cancer patients was constructed using these stratification and survival time data (Supplementary Fig. 3). The model's effectiveness was evaluated using a time-dependent ROC curve. Kaplan–Meier survival analysis revealed a greater likelihood of poor prognosis in the high-risk group. Consequently, the experimental group model demonstrated superior performance (1-year AUC: 0.692, 2-year AUC: 0.699, 4-year AUC: 0.653, 8-year AUC: 0.591) compared with the control group model (1-year AUC: 0.678, 2-year AUC: 0.697, 4-year AUC: 0.633, 8-year AUC: 0.602) as depicted in Fig. 6.

To further validate the advantages of the LCBD in constructing prognostic models for lung cancer biomarkers, we conducted a systematic comparison with other databases. Since the LCcircDB database, developed by Dr. Xia in 2018, remains inaccessible, and the LCMD database contains only lung cancer metabolite biomarkers incompatible with the transcriptomic nature of the TCGA-LUAD bulk RNA-seq dataset, we could not compare these with databases such as MethMarkerDB, MarkerDB, and TheMarker, which encompass all types of lung cancer biomarkers. Specifically, we built a prognostic model by integrating the differentially expressed genes identified from the TCGA-LUAD dataset with biomarkers from the LCBD. Kaplan–Meier survival analysis was then performed to compare our model with those derived from the MethMarkerDB, MarkerDB, and TheMarker databases, thereby comprehensively assessing the

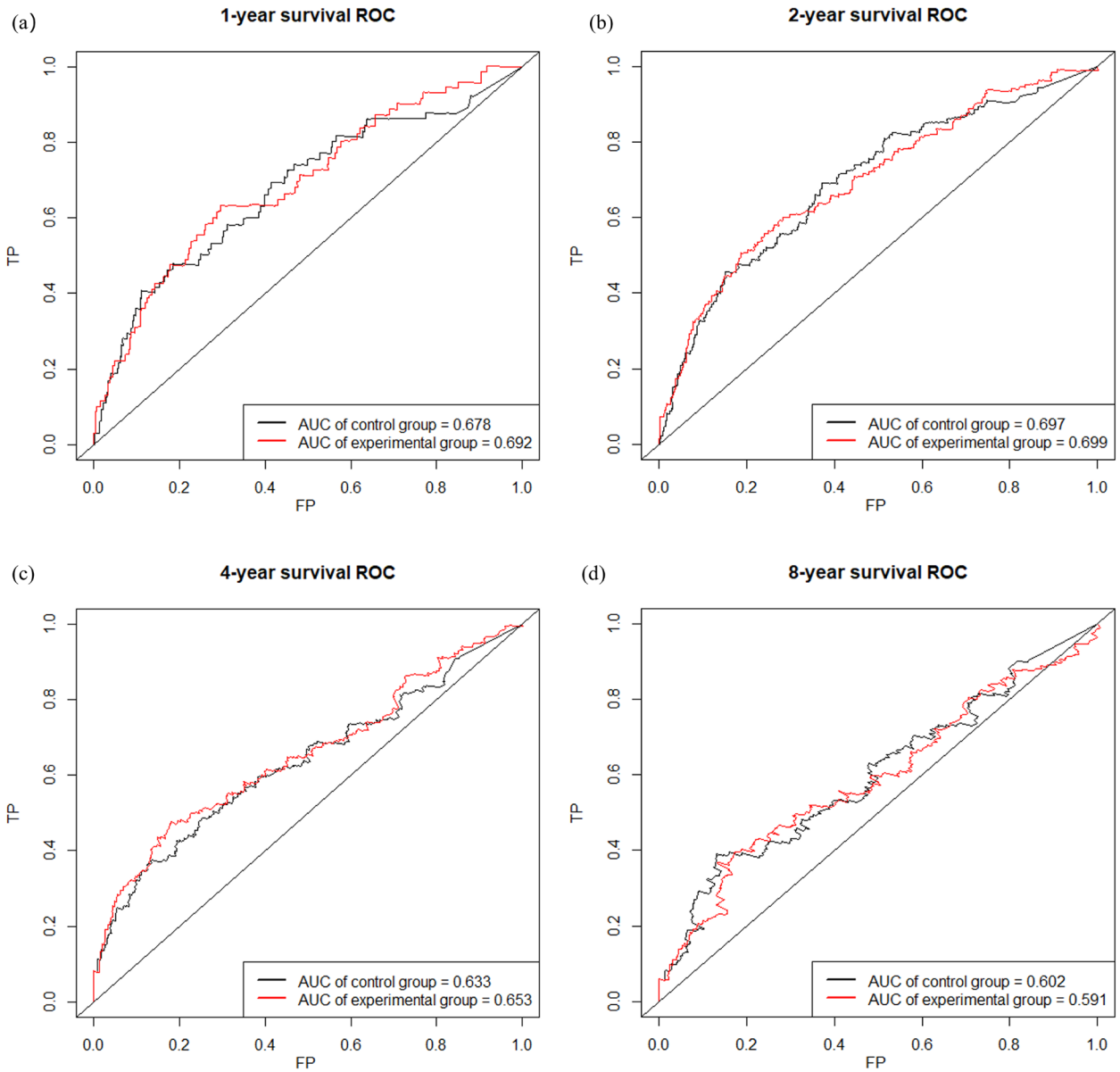


Fig. 6 Time-dependent ROC curves comparing prognostic models derived from the LCBD with those from other databases. **a** ROC curve for 1-year survival, **b** ROC curve for 2-year survival, **c** ROC curve for 4-year survival, and (**d**) ROC curve for 8-year survival. The ROC curve of the experimental group is shown in orange, and the ROC curve of the control group is shown in red

superior predictive performance and clinical applicability of LCBD. The experimental results indicate that, with respect to the AUC scores for 1-year, 2-year, 4-year, and 8-year survival, the LCBD-based model outperformed the models constructed using the other databases. Specifically, the LCBD-based model achieved AUCs of 0.75, 0.737, 0.747, and 0.737, respectively, which were significantly higher than those of the MethMarkerDB-based model (0.743, 0.732, 0.723, and 0.719) and the MarkerDB-based model (0.743, 0.73, 0.723, and 0.724), while the TheMarker-based model yielded comparatively lower AUCs of 0.668, 0.664, 0.637, and 0.637 (Fig. 7). Furthermore, to rigorously substantiate the statistical superiority of the LCBD database in prognostic modeling, we computed the AUC for prognostic models derived from LCBD and three comparator databases (MarkerDB, MethMarkerDB, and TheMarker) via fivefold cross-validation. Leveraging this cross-validation data, we performed t-tests to assess intergroup differences in AUC between LCBD and the comparator databases. The

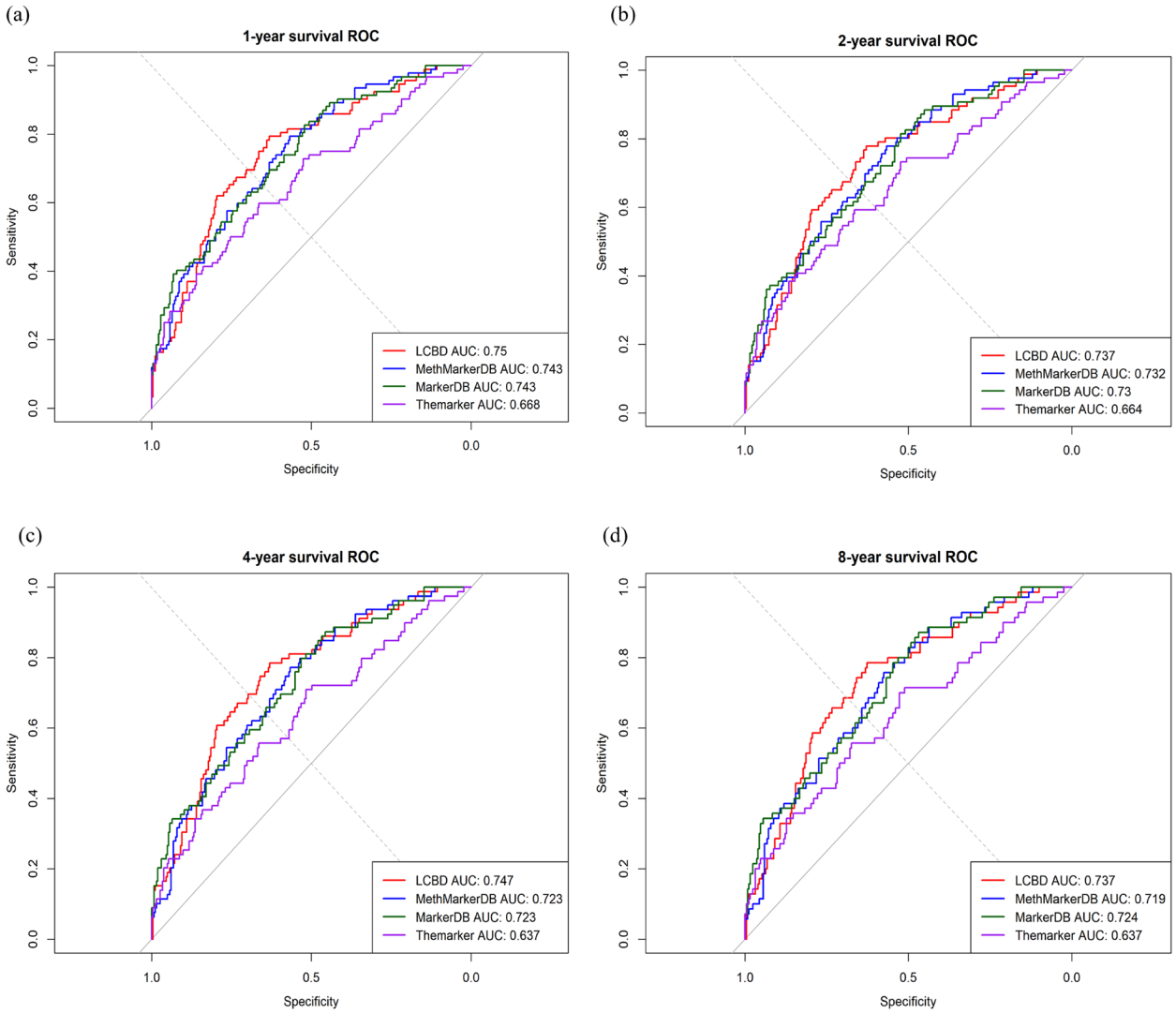


Fig. 7 Time-dependent ROC curves comparing prognostic models derived from the LCBD with those from other databases. **a** ROC curve for 1-year survival; **b** ROC curve for 2-year survival; **c** ROC curve for 4-year survival; **d** ROC curve for 8-year survival. The red, blue, green, and purple curves represent the LCBD, MethMarkerDB, MarkerDB, and TheMarker groups, respectively

findings revealed that, in the majority of comparisons, LCBD exhibited significantly higher AUC values than the other databases ($p < 0.05$), thereby confirming its enhanced performance in prognostic modeling (see Supplementary Table 4 for details).

In summary, biomarkers from the LCBD can significantly contribute to the refinement and accuracy of prognostic prediction models in lung cancer research.

Diagnostic model for lung cancer

In the second case study, we utilized biomarkers from the LCBD to develop a diagnostic model for lung cancer patients. Clinical information and gene expression profiles for LUAD patients were sourced from The Cancer Genome Atlas (TCGA-LUAD), and a control group was established to reduce the influence of irrelevant variables. The experimental group included DEGs from both the LCBD and the original LUAD dataset (a total of 1380 genes), and the control group consisted of only the original LUAD DEGs (519 genes). For the proteomics diagnostic model, we defined the DEPs found in both the LCBD and the original LUAD dataset as the experimental group (6,385 proteins), and DEPs identified exclusively in the original LUAD dataset (6,225 proteins) were designated as the control group. We identified lung cancer-related genes/proteins from both datasets by calculating the KL divergence and selected the 10 genes/proteins with the greatest KL divergence f scores as inputs for the

diagnostic model. A deep neural network with two hidden layers was then constructed to build the diagnostic model.

The results in the TCGA dataset revealed that the model achieved an accuracy of 94.65% and an area under the curve (AUC) value of 0.9179 in the experimental group. These metrics greatly exceeded the diagnostic performance of the model in the control group, which had an accuracy of 92.59% and an AUC of 0.8141 (Fig. 8a). Moreover, the results for the CPTAC dataset reveal that the model achieved an accuracy of 83.72% and an AUC (area under the curve) value of 0.9481 in the control group. In comparison, the model exhibited improved diagnostic performance in the

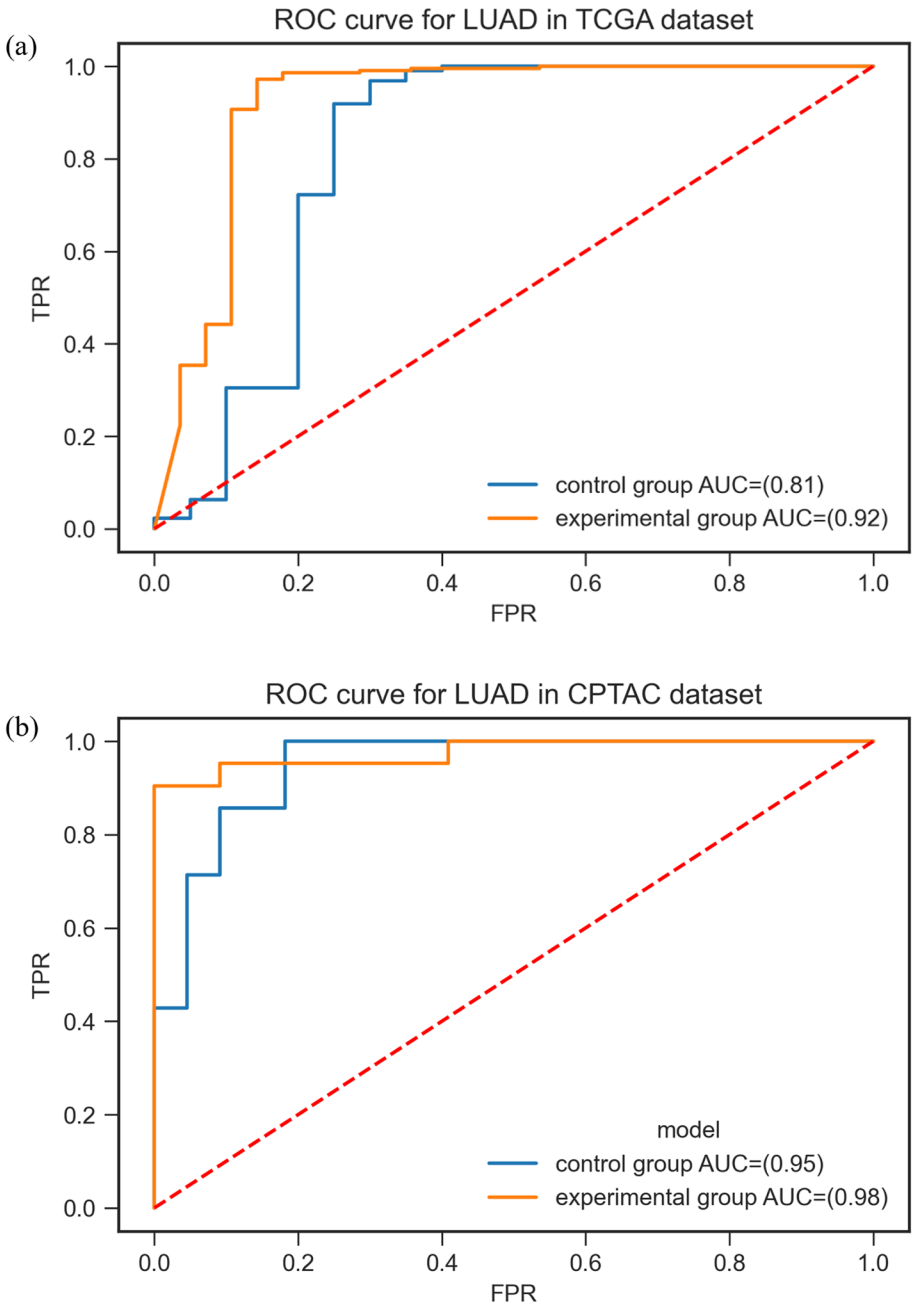


Fig. 8 ROC curves for the LUAD diagnostic model. **a** ROC curve of the LUAD diagnostic model derived from the TCGA dataset; **b** ROC curve of the LUAD diagnostic model derived from the CPTAC dataset. The horizontal axis represents the false-positive rate, and the vertical axis represents the true positive rate, with the area under the ROC curve (AUC) representing model performance. The ROC curve of the experimental group containing biomarkers from the LCBD is shown in blue, and the ROC curve of the control group without biomarkers from the LCBD is shown in orange

experimental group, with an accuracy of 90.70% and an AUC of 0.9751 (Fig. 8b). When validated with multi-omics data, biomarkers from the LCBD can be used to construct a more accurate lung cancer diagnostic model.

To further demonstrate the advantages of our constructed LCBD, we developed a diagnostic prediction model based on KL divergence and compared its performance with that of other databases, including MarkerDB, TheMarker, and MethMarkerDB. The results indicate that the diagnostic accuracy of the TheMarker database was 95.06% with an AUC of 0.93; the MarkerDB database achieved an accuracy of 93.00% with an AUC of 0.97; and the MethMarkerDB database attained an accuracy of 93.83% with an AUC of 0.94. In contrast, the LCBD exhibited improved diagnostic performance, with an accuracy of 98.77% and an AUC of 0.99 (as shown in Fig. 9). Furthermore, to validate the statistical superiority of the LCBD database in diagnostic modeling, we computed the area under the AUC for diagnostic models derived from LCBD and other databases (MarkerDB, MethMarkerDB, and TheMarker) via fivefold cross-validation, followed by t-tests to evaluate intergroup differences in AUC. The results indicated that, in the majority of comparisons, the AUC for LCBD was significantly higher than that of the comparator databases ($p < 0.05$), confirming its pronounced performance advantage in diagnostic modeling (see Supplementary Table 5 for details).

Immune infiltration model for lung cancer

Using the gene signatures of 28 immune cell types reported by Jia Q [37], the enrichment scores for these immune cells in both the training and control sets were calculated with the "GSVA" package (v1.42.0) [34] and the single-sample gene set enrichment analysis (ssGSEA) method. The matrix of immune cell enrichment scores was subsequently normalized and categorized using the "Non-negative Matrix Factorization" (NMF) software package. The NMF analysis was conducted with the rank varying from 2 to 6 and the number of runs (nrun) set to 30. After cogene typing indices across different rank

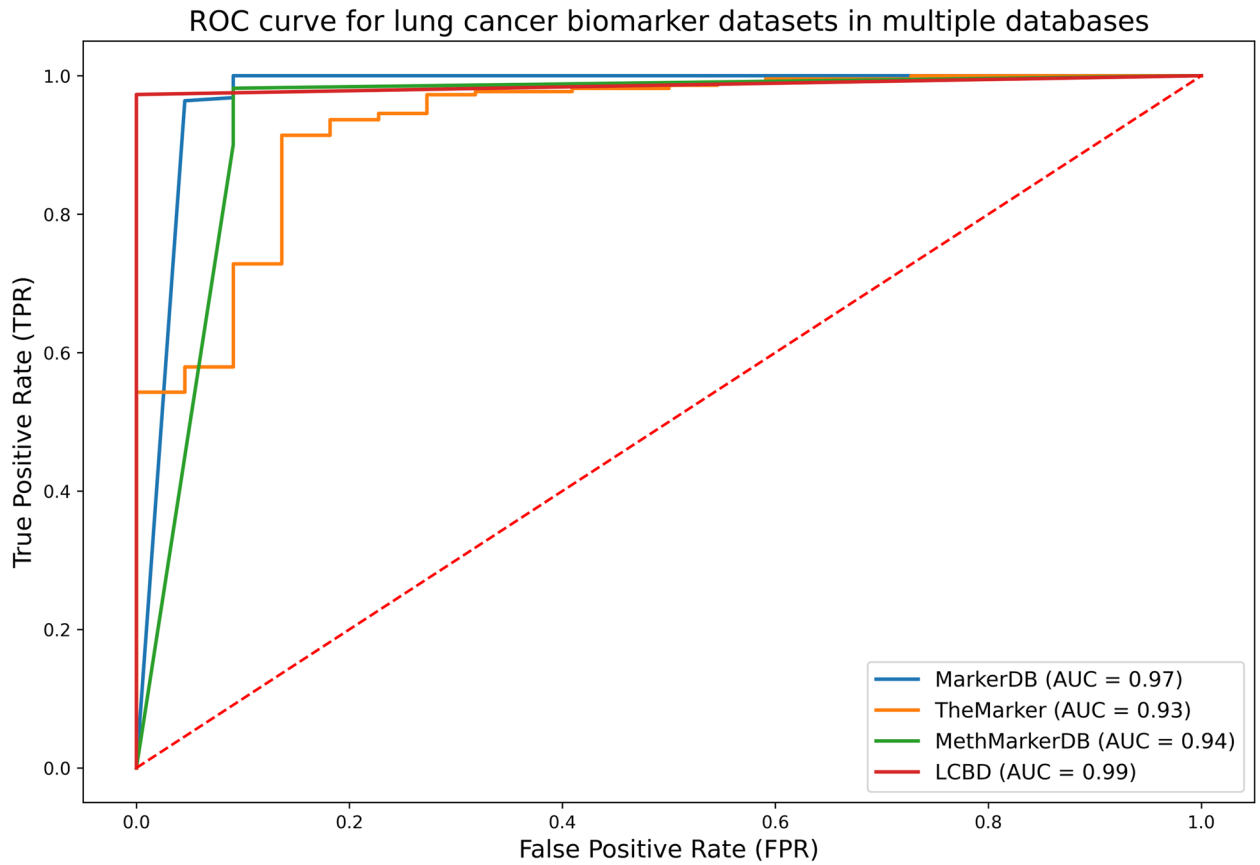


Fig. 9 ROC curves for lung cancer biomarker datasets from multiple databases. The horizontal axis represents the false positive rate, and the vertical axis represents the true positive rate, with the area under the ROC curve (AUC) reflecting model performance. The blue, orange, green, and red curves represent the ROC curves for biomarkers from the MarkerDB, TheMarker, MethMarkerDB, and LCBDs, respectively

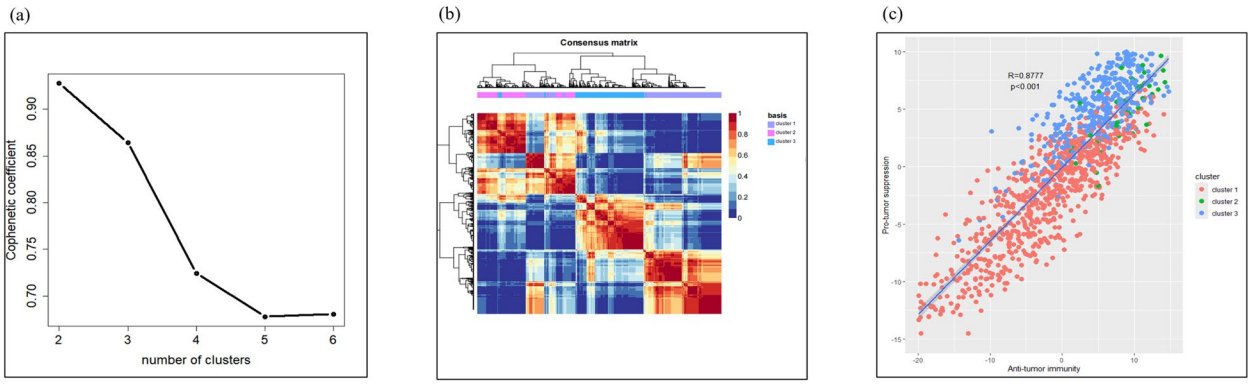


Fig. 10 Changes in the cophenetic index with the number of clusters, a heatmap of NMF tumour sample classification at rank = 3, and correlation analysis of NMF classification results. **a** On the basis of the graph showing changes in the cophenetic index relative to cluster numbers, the samples are classified into three groups: immune desert (Cluster 1), immune exclusion (Cluster 2), and immune inflammatory (Cluster 3). **b** NMF classification was conducted using an enrichment score matrix of 28 immune cells in the samples, and the resulting heatmap is displayed. **c** Correlation analysis of the NMF classification results was conducted on the basis of the immune enrichment scores

values were evaluated (Fig. 10a), a rank of 3 was selected, leading to the classification of patients into three clusters: immune desert (Cluster 1), immune exclusion (Cluster 2), and immune inflammatory (Cluster 3). A heatmap depicting the sample types was also generated (Fig. 10b). However, a subsequent correlation analysis between immuno-enrichment scores and NMF typing (Fig. 10c) revealed that Clusters 2 and 3 presented greater correlations that were concentrated at the upper end of the correlation curve. Considering the smaller sample size of Cluster 2 and its similarity in characteristics to Cluster 3, these two groups were amalgamated into a new subtype, termed the high immune infiltration type (C2), whereas the original Cluster 1 was designated as the low immune infiltration type (C1).

We used KNN [35] and random forest [36] classifiers to compare the prediction performance of immune infiltration typing between the control group and the experimental group. The ROC curve analysis based on the KNN results revealed that the model had an AUC of 0.86 in the experimental group, which was better than that in the control group (AUC, 0.8) (Fig. 11). The models based on the random forest classifier yielded similar results (experimental group: AUC, 0.94; control group: AUC, 0.92) (Fig. 11). These results indicate that the immune infiltration model incorporating biomarkers from the LCBD has superior predictive performance compared with the control model, confirming the reliability and effectiveness of the biomarker data in the LCBD. Furthermore, biomarkers in the LCBD have significant predictive value for determining the immune infiltration status of lung cancer patients.

To further validate the advantages of our constructed LCBD in predicting immune infiltration subtypes, we evaluated this task using both KNN and Random Forest models, and compared the results with those obtained from corresponding models based on the MarkerDB, TheMarker, and MethMarkerDB databases. ROC curve analysis of the Random Forest model (Fig. 12) revealed that the AUC for LCBD reached 0.9383, which is significantly higher than those for the other databases (MarkerDB: 0.9178; MethMarkerDB: 0.9183; TheMarker: 0.9369). Similarly, the KNN model produced comparable results (LCBD: 0.8831; MarkerDB: 0.8858; MethMarkerDB: 0.8710; TheMarker: 0.8024), further demonstrating that LCBD exhibits superior accuracy and robustness in predicting immune infiltration subtypes. Furthermore, to rigorously validate the statistical superiority of the LCBD database in immune infiltration modeling, we computed the AUC for immune infiltration models derived from LCBD and comparator databases (MarkerDB, MethMarkerDB, and TheMarker) using fivefold cross-validation. Based on the cross-validation data, we conducted t-tests to assess intergroup differences in AUC. The findings revealed that the AUC for LCBD was significantly higher than that of the comparator databases ($p < 0.05$), thereby confirming its pronounced performance advantage in constructing immune infiltration models (see Supplementary Table 6 for details).

Clinical applications and challenges of LCBD biomarkers

The LCBD is a comprehensive platform designed to advance early detection and personalized treatment of lung cancer. By integrating biomarker data from multiple studies, including clinically validated retrospective and prospective research, LCBD enhances research efficiency, minimizes experimental redundancy, and reduces the costs associated with preliminary screening. In early lung

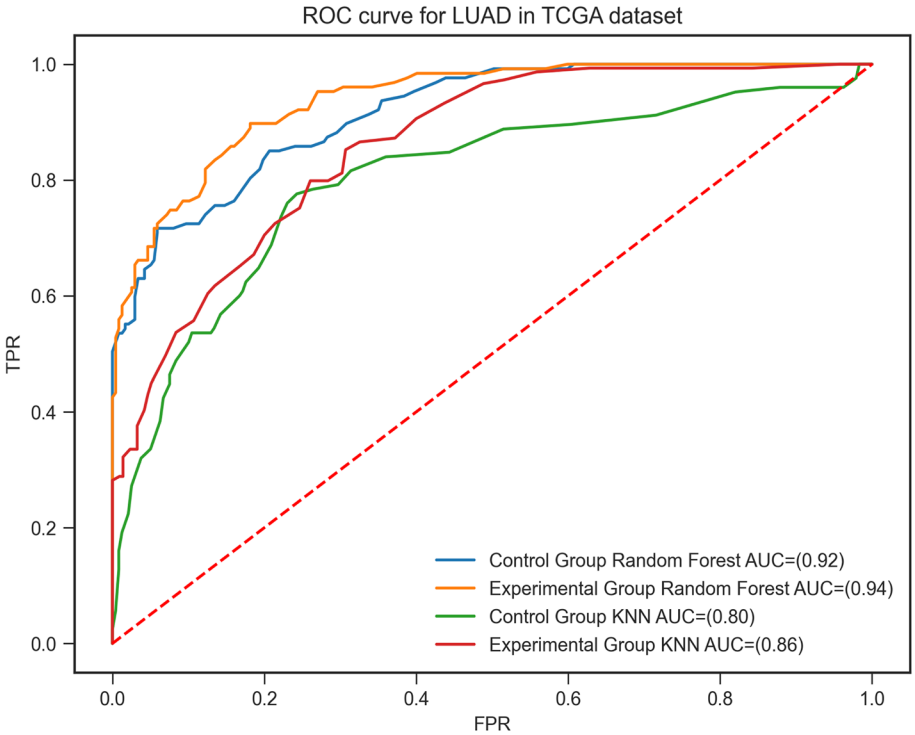


Fig. 11 ROC curves for the control and experimental groups based on the KNN and random forest methods. The horizontal axis represents the false-positive rate, and the vertical axis represents the true-positive rate, with the area under the ROC curve (AUC) representing model performance. The ROC curves of the control group (without LCBD biomarkers) using the random forest model and the KNN model are shown in blue, and the ROC curves of the experimental group (with LCBD biomarkers) using the random forest model and the KNN model are shown in orange

cancer detection, the LCBD systematically identifies biomarker combinations with the highest diagnostic value. Physicians can utilize LCBD data, liquid biopsy results, and risk prediction models to detect high-risk individuals at an earlier stage, thereby improving early diagnosis rates, increasing patient survival, and reducing the incidence of late-stage lung cancer. For precision medicine, the LCBD provides a searchable platform that allows clinicians to retrieve immunotherapy-predictive biomarkers. The LCBD also facilitates treatment selection on the basis of genetic mutations and molecular profiles, including PD-L1 expression levels, tumour mutational burden (TMB), and microsatellite instability (MSI), optimizing immunotherapy and targeted therapy decisions.

However, translating biomarkers into clinical applications poses challenges, particularly in integrating these biomarkers into existing diagnostic workflows. Biomarkers curated in the LCBD align with modern detection technologies, such as next-generation sequencing (NGS), liquid biopsy, and proteomics analyses, ensuring their feasibility in clinical settings. Collaboration among pathologists, bioinformaticians, and clinicians is essential to bridge the gap between theoretical potential value and real-world application. Furthermore, regulatory approval remains a major hurdle. Different applications—screening, diagnosis, prognosis assessment, and companion diagnostics (CDx)—require compliance with distinct standards. The LCBD employs a tiered classification system, categorizing biomarkers into experimental validation, clinical trials, and regulatory approval and providing a clear regulatory status to aid in clinical decision-making. Despite its contributions to accelerating lung cancer biomarker translation, the process remains a complex and ongoing challenge.

LCBD limitations and future plans

As a literature-derived database focused on lung cancer biomarkers, the LCBD primarily aggregates published research evidence. However, the inherent publication bias in scientific literature, with negative or statistically nonsignificant findings being underrepresented, risks inflating the perceived clinical utility of certain biomarkers during data curation.

The landscape of lung cancer biomarkers is rapidly evolving, driven by multiple factors: (1) commercialization of novel detection platforms (e.g., multiomics-based assays), (2) regulatory approvals of diagnostic techniques (EMA/FDA-cleared), (3)

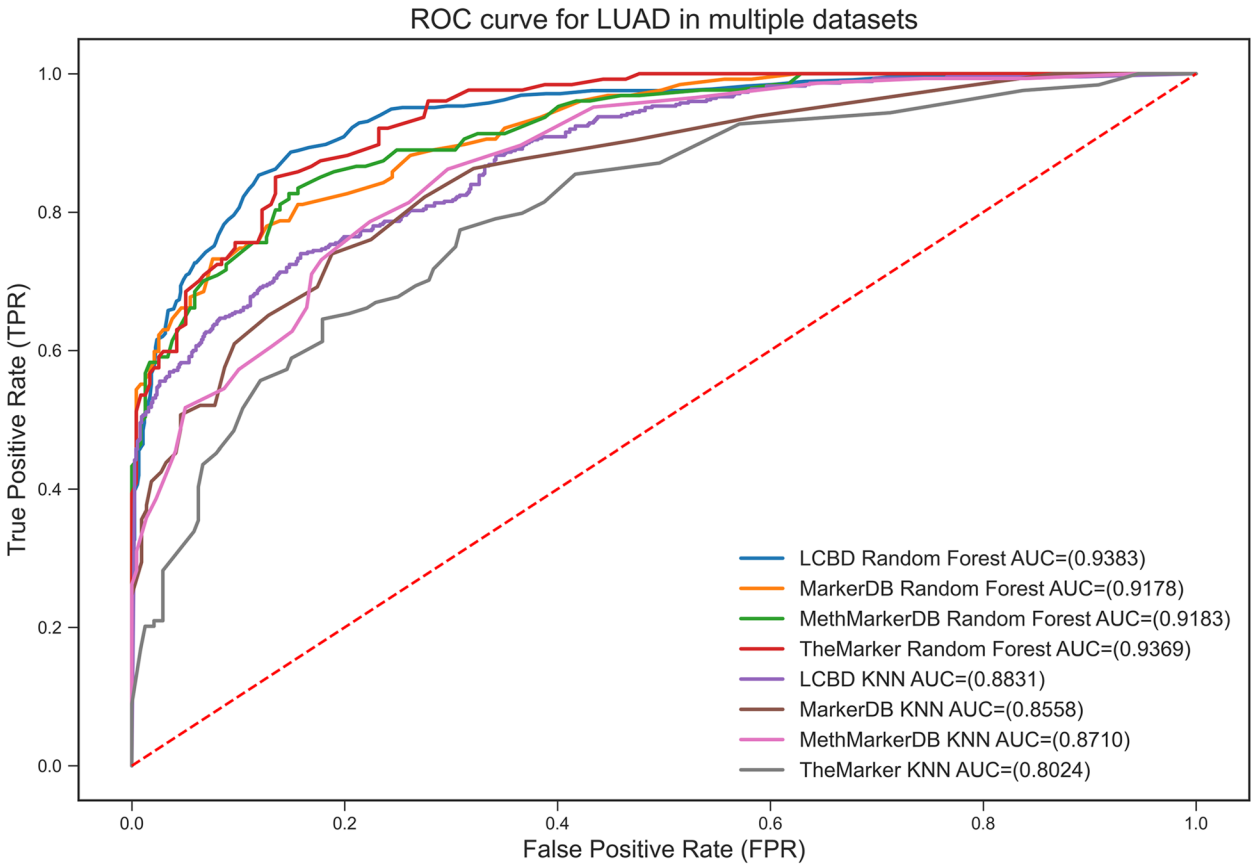


Fig. 12 ROC curves for lung cancer biomarker datasets in multiple databases based on KNN and Random Forest. The blue, orange, green, and purple curves represent the ROC curves of the LCBD, MarkerDB, MethMarkerDB, and TheMarker databases, respectively, using the Random Forest model. The pink, brown, olive green, and gray curves represent the ROC curves of the LCBD, MarkerDB, MethMarkerDB, and TheMarker databases, respectively, using the KNN model

advancements in laboratory-developed tests (LDTs), and (4) continuous publication and patenting of biomarker discoveries. Importantly, the clinical applicability of biomarkers requires dynamic re-evaluation as emerging evidence may change their diagnostic/prognostic validity. These dynamics necessitate systematic updates of the LCBD to maintain its translational relevance.

To address these challenges, we plan to implement scheduled updates every 2–3 years with the following enhancements: (1) integration of preprint repositories (e.g., bioRxiv, medRxiv) to capture biomarker candidates prior to journal-driven selection bias, thereby identifying high-potential markers overlooked by traditional publication filters; (2) inclusion of negative studies and conflicting evidence reports (e.g., biomarkers validated in specific cohorts but invalidated in others) to counterbalance publication bias; and (3) algorithm-driven selection criteria to prioritize biomarkers with cross-platform reproducibility and multicentre validation records, reducing reliance on single-study claims. Furthermore, we plan to introduce two new modules to the LCBD platform: a diagnostic model module and a prognostic model module. These modules will allow users to upload patient-related data and utilize the platform's integrated diagnostic or prognostic models for analysis, thereby enabling diagnostic and prognostic predictions. We expect that periodic updates and iterations of LCBD will substantially improve its clinical utility, delivering more precise lung cancer biomarker services to researchers and clinicians.

Conclusions

In this study, we developed the Lung Cancer Biomarker Database (LCBD), a comprehensive and integrated repository tailored for lung cancer biomarkers. The establishment of the LCBD provides researchers and clinicians with convenient access to extensive data on a wide array of lung cancer biomarkers, including genes, proteins, miRNAs, lncRNAs, circRNAs, and metabolites. This database not only facilitates the retrieval of specific lung cancer-related biomarkers but also enriches research into disease pathogenesis, playing a pivotal role

in early diagnosis, disease classification, and prognosis evaluation.

The creation of the LCBD has significantly driven research and clinical applications in the lung cancer biomarker field, providing the global lung cancer research community with an integrated and efficient platform. This platform improves the allocation of existing research resources and underpins the innovation and development of strategies for the prevention, diagnosis, and treatment of lung cancer. Consequently, we believe that the LCBD will become an essential component of future lung cancer research and treatment strategies, improving the management and outcomes of the disease.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13883-w.

Supplementary Material 1.

Acknowledgements
We are deeply grateful to Jin Zhang and Zunian Wang for their expert technical support during the development of our website.

Authors' contributions
Conceptualization, Y.L., B.L., and K.S.; Data Collection, Z.T., Y.Y., Y.W., L.W., Y.X., and Y.L.; Methodology, Z.T., Y.X., B.L; Case Studies Construction, L.W., M.B; Writing – original draft, Y.L. and Z.T.; Writing – review & editing, Y.L., B.L. and K.S.; Funding acquisition, Y.L. and B.L.; Supervision, K.S.

Funding
This work was supported by the National Natural Science Foundation of China (Grant No. 62101087), the China Postdoctoral Science Foundation (Grant No. 2021MD703942), the Chongqing Municipal Postdoctoral Science Special Foundation (Grant No. 2021XM2016), the Science and Technology Research Program of the Chongqing Municipal Education Commission (Grant No. KJQN202100642, KJQN202100538), the Chongqing Natural Science Foundation (cstc2021jcyj-msxmX0834) and the Innovation and Entrepreneurship Training Program for College Students in Chongqing (Grant No. S202010617010).

Data availability
The primary website for the LCBD is http://lcbd.biomarkerdb.com, with an alternative mirror site available at http://lcbd.lyhbio.com. The datasets used in this study are available from the TCGA repository (https://portal.gdc.cancer.gov/projects/TCGA-LUAD) and the CPTAC database, which was accessed via LinkedOmics (https://www.linkedomics.org/data_download/CPTAC-LUAD/).

Declarations

Ethics approval and consent to participate
This study utilized publicly available datasets and did not involve any experimental procedures on humans or animals. Therefore, specific ethics approval and consent to participate were not required for this secondary data analysis. The use of these datasets complies with the terms and conditions set by the data providers, and the analysis adhered to all relevant ethical guidelines.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 June 2024 Accepted: 7 March 2025
Published online: 15 March 2025

References

1. Bade BC, Dela Cruz CS. Lung Cancer 2020: Epidemiology, Etiology, and Prevention. Clin Chest Med. 2020;41(1):1–24.
2. Siegel RL, Giaquinto AN. Cancer statistics, 2024. CA: a cancer journal for clinicians. 2024;74(1):12–49.
3. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023;73(1):17–48.
4. Kuhn E, Morbini P, Cancellieri A, Damiani S, Cavazza A, Comin CE. Adenocarcinoma classification: patterns and prognosis. Pathologica. 2018;110(1):5–11.
5. Cokkinides V, Albano J, Samuels A, Ward M, Thum J. American cancer society: Cancer facts and figures. Atlanta: Am Can Soc. 2005;2017:4–8.
6. Hu J, Qian G-S, Bai C-X. Society LCSGotCT, Group tCAALCE: Chinese consensus on early diagnosis of primary lung cancer (2014 version). Cancer. 2015;121(S17):3157–64.
7. Callejon-Leblic B, García-Barrera T, Pereira-Vega A, Gómez-Ariza JL. Metabolomic study of serum, urine and bronchoalveolar lavage fluid based on gas chromatography mass spectrometry to delve into the pathology of lung cancer. J Pharm Biomed Anal. 2019;163:122–9.
8. Ning J, Ge T, Jiang M, Jia K, Wang L, Li W, Chen B, Liu Y, Wang H, Zhao S, et al. Early diagnosis of lung cancer: which is the optimal choice? Aging. 2021;13(4):6214–27.
9. García-Giménez JL, Seco-Cervera M, Tollefsbol TO, Romá-Mateo C, Peiró-Chova L, Lapunzina P, Pallardó FV. Epigenetic biomarkers: Current strategies and future challenges for their use in the clinical laboratory. Crit Rev Clin Lab Sci. 2017;54(7–8):529–50.
10. Calvayrac O, Pradines A, Pons E, Mazières J, Guibert N. Molecular biomarkers for lung adenocarcinoma. Eur Respir J. 2017;49(4):1601734.
11. Wu WS, Wu HY, Wang PH, Chen TY, Chen KR, Chang CW, Lee DE, Lin BH. Chang WC-W, Liao PC. LCMD: lung cancer metabolome database. Comput Struct Biotechnol J. 2022;20:65–78.
12. Xia Y, Chen Z. Lung cancer circRNA bioinformatics analysis and database development. Wuhan: Huazhong Agricultural University; 2018.
13. Zhang Y, Zhou Y, Zhou Y, Yu X, Shen X, Hong Y, Zhang Y, Wang S, Mou M. TheMarker: a comprehensive database of therapeutic biomarkers. Nucleic Acids Res. 2024;52(D1):D1450–64.
14. Zhu Z, Zhou Q, Sun Y, Lai F, Wang Z, Hao Z, Li G. MethMarkerDB: a comprehensive cancer DNA methylation biomarker database. Nucleic Acids Res. 2024;52(D1):D1380–92.
15. Wishart DS, Bartok B, Oler E, Liang KYH, Budinski Z, Berjanskii M, Guo A, Cao X, Wilson M. MarkerDB: an online database of molecular biomarkers. Nucleic Acids Res. 2021;49(D1):D1259–67.
16. Consortium TU, Bateman A, Martin M-J, Orchard S, Magrane M, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bye-A-Jee H, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–31.
17. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155–62.
18. Volders P-J, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, Vandesompele J. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. 2019;47(D1):D135–9.
19. Liu M, Wang Q, Shen J, Yang BB, Ding X. Circbank: a comprehensive database for circRNA with standard nomenclature. RNA Biol. 2019;16(7):899–905.
20. Pian C, Zhang G, Tu T, Ma X, Li F. LncCeRBase: a database of experimentally validated human competing endogenous long non-coding RNAs. Database (Oxford). 2018;2018:bay061.
21. Huang HY, Lin YCD, Li J, Huang KY, Shrestha S, Hong HC, Tang Y, Chen YG, Jin CN, Yu Y, et al. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. Nucleic Acids Res. 2019;48(D1):D148-D154.
22. Zheng Y, Luo H, Teng X, Hao X, Yan X, Tang Y, Zhang W, Wang Y, Zhang P, Li Y, et al. NPInter v5.0: ncRNA interaction database in a new era. Nucleic Acids Res. 2023;51(D1):D232-D239.
23. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al. DrugBank 5.0: a major

update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074-D1082.
24. Chen X. TTD: Therapeutic Target Database. Nucleic Acids Res. 2002;30(1):412–5.
25. Mukhtar H, Kang-Myo K, Chaudhry SA, Akbar AH, Ki-Hyung K, Yoo SW. LNMP- Management architecture for IPv6 based low-power wireless Personal Area Networks (6LoWPAN). Piscataway: IEEE. In: NOMS 2008 - 2008 IEEE Network Operations and Management Symposium. 2008:417–424.
26. Gormley C, Tong Z. Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. Sebastopol: "O'Reilly Media, Inc."; 2015.
27. Burns B, Grant B, Oppenheimer D, Brewer E, Wilkes J. Borg, Omega, and Kubernetes: Lessons learned from three container-management systems over a decade. Queue. 2016;14(1):70–93.
28. Mj G, B C, M H, K R, F M, A K, A B, Y L, D R, An B, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechno. 2020;38(6):675–678.
29. Zhou J, Wang X, Li Z, Jiang R. Construction and Validation of Prognostic Risk Score Model of Autophagy Related Genes in Lung Adenocarcinoma. Chin J Lung Cancer. 2021;24(8):557–66.
30. Liu S, Yao W. Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection. BMC Bioinformatics. 2022;23(1):175.
31. Yang L, Wei S, Zhang J, Hu Q, Hu W, Cao M, Zhang L, Wang Y, Wang P, Wang K. Construction of a predictive model for immunotherapy efficacy in lung squamous cell carcinoma based on the degree of tumor-infiltrating immune cells and molecular typing. J Transl Med. 2022;20(1):364.
32. Kullback S, Leibler RA. On Information and Sufficiency. Ann Math Stat. 1951;22(1):79–86.
33. Liu W, Wang T, Chen S, Tang A. Hierarchical Clustering of Gene Expression Data with Divergence Measure. 2009 3rd International Conference on Bioinformatics and Biomedical Engineering. Piscataway: IEEE. 2009:1–3.
34. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
35. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3–7, 2003 Proceedings: 2003: Springer; 2003:986–996.
36. Breiman L. Random forests. Machine learning. 2001;45:5–32.
37. Jia Q, Wu W, Wang Y, Alexander PB, Sun C, Gong Z, Cheng J-N, Sun H, Guan Y, Xia X, et al. Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. Nat Commun. 2018;9(1):5361.

Publisher's Note