scientific data

Check for updates

OPEN

DATA DESCRIPTOR

SexTumorDB: a comprehensive resource of sex-dependent tumor landscape at single-cell resolution

Ruya Sun[1,4] ✉, Qiqi Deng[1,4] & Di Wang[1,2,3] ✉

Sex is one of the most fundamental biological variables, with a profound impact on tumor incidence, mortality, treatment response and prognosis, making it a critical consideration for precision oncology. Despite its well-established role in tumors, sex bias is often overlooked in both basic research and clinical practice. Our understanding of sex disparities in tumors, especially in non-reproductive tumors and non-malignant components, remains limited, necessitating a comprehensive investigation of sexual dimorphism in cancer. To address this gap, we developed the SexTumorDB database, an integrated resource containing RNA sequencing profiles of non-reproductive tumors at single-cell resolution. SexTumorDB covers 13 cancer types with the highest incidence in either males or females, encompassing 2,014,043 cells from 532 samples, including 319 male and 213 female samples. The SexTumorDB datasets were rigorously curated, processed, and standardized. By making this resource publicly available through online repositories and web applications, we aim to advance the understanding of sex disparities in the tumor microenvironments (TMEs) and foster the development of sex-specific anticancer treatment strategies in the future.

Background & Summary

Sex, as determined by chromosomes, is a fundamental biological variable for population stratification[1]. Sex disparities shape distinct health and disease spectrums between females and males, exhibiting an overall male predominance in cancer incidence and mortality[2,3]. According to the most recent global cancer statistics, the male-to-female cancer incidence ratio can reach 7.78 (larynx cancer), and the mortality ratio can be twice as high in males (e.g., esophageal carcinoma, stomach cancer)[4,5]. Paradoxically, females present poorer responsiveness to immunotherapy, which might be partially explained by the accumulation of worse driver mutations[6]. The sexual dimorphisms underlying cancer susceptibility and outcome highlight the need for sex-specific antitumor treatment strategies[7]. While in both basic and translational research, sex bias is rarely emphasized, particularly in studies of tumors arising in non-reproductive organs. Current knowledge of the mechanistic basis for sexual dimorphism in cancer still remains limited, necessitating comprehensive investigations of sexual dimorphism across tumor microenvironments (TMEs) and identification of sex-specific therapeutic targets.

With the increased attention to sex bias in diseases, a few efforts have been made to decode the sexual dimorphism in cancer. However, due to the limitations in many aspects, including sample availability, cohort size and sequencing technology, most of these studies remain at the level of whole tumor tissue or mainly focus on malignant cells[8–11], systematic analyses of sex disparities within the intricate TME ecosystems are missing. In recent years, the rapidly developing single-cell RNA sequencing technology brings us the opportunity to gain deep insights into the complex cell components within the TME and their corresponding functional states, making it possible to explore the mechanistic basis underlying various tumors at an unprecedented resolution[12]. A series of tumor microenvironment landscapes have been constructed based on the single-cell sequencing profiles of numerous patients, revealing the inter-tumor as well as intra-tumor heterogeneity across cancer cohorts[13–25]. These studies provide valuable data support to summarize the sex disparities across various TMEs from a broad

[1]Department of Orthopaedic Surgery of Sir Run Run Shaw Hospital, and Institute of Immunology, Zhejiang University School of Medicine, Hangzhou, 310058, China. [2]Department of Colorectal Surgery and Oncology, Key Laboratory of Cancer Prevention and Intervention, Ministry of Education, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, 310058, China. [3]Liangzhu Laboratory, Zhejiang University Medical Center, Hangzhou, 311113, China. [4]These authors contributed equally: Ruya Sun, Qiqi Deng. ✉e-mail: sunruya@zju.edu.cn; diwang@zju.edu.cn

perspective. However, due to the differences in study scope and research conditions, these datasets vary a lot in sample collection and sequencing procedures, bringing great difficulty to conduct integrative analyses of these data, especially to researchers without bioinformatics expertise. Therefore, there is an urgent need to integrate the published sequencing profiles of cancer cohorts and compile a comprehensive resource of the sex-dependent tumor landscape to facilitate the systematic investigation of sexual dimorphism in cancer.

In this study, through rigorous literature screening, data curation and processing, cell type annotation, data standardization, and integration, we constructed the SexTumorDB database, a comprehensive landscape of sex-dependent tumor microenvironments at single-cell resolution, with detailed cell type annotation and sex classification. This integrated data source includes all 13 non-reproductive system tumors with the top 10 incidence rates in either females or males, encompassing a total of 2,014,043 single cells from 532 samples. The SexTumorDB comprises 33 major cell types that belong to the tumor, immune, and stromal cell components within various tumor microenvironments. To facilitate high-quality investigations, all included samples are from the primary site of corresponding tumors and are from treatment-naïve donors. Among them, 74.62% of cells are from tumor samples (362 samples), including 139 female samples and 223 male samples, the remaining 25.38% cells are from healthy donors or normal tissues adjacent to tumors (170 samples). To enable easy access to and convenient reuse of these datasets, we upload all relevant data in the publicly accessible Zenodo repository[26] (https://doi.org/10.5281/zenodo.16742263), and provide web applications for relevant researchers to quickly explore the sex disparities across diverse tumor microenvironments. Collectively, the SexTumorDB provides basic and reliable data support for studies on sex disparities in cancer, fostering the mechanistic understanding of sexual dimorphism and the development of sex-specific anticancer treatment strategies.

Methods

The SexTumorDB database is developed to provide researchers with a user-friendly, public data source to investigate sex-dependent disparities across diverse tumor microenvironments (TMEs). A standard workflow was applied for data curation and processing, and the workflow is summarized below (Fig. 1).

Data collection. To ensure comprehensive data coverage in the SexTumorDB database, we first defined the scope of target cancer types. The most recent global cancer statistics for both female and male populations (Globocan 2022, version 1.1) were obtained from the WHO website (https://gco.iarc.who.int/today)[4]. Non-reproductive cancers with the age-standardized incidence ranking among the top 10 in either sex were selected, and their union was chosen as the scope of data collection.

Data collection was then conducted manually in public resources, primarily PubMed (www.ncbi.nlm.nih.gov/pubmed/) and the GEO (www.ncbi.nlm.nih.gov/geo/) platform. For each cancer type, we searched for datasets that met the following criteria: (1) RNA-seq profile of cancer cohort at single-cell resolution; (2) inclusion of sex classification for sample donors; (3) samples taken from the primary site of the corresponding cancer type; (4) donors were treatment-naïve; (5) cell components were not sorted. Datasets meeting all these criteria were evaluated based on sample size and data accessibility. Also, to mitigate batch effects arising from different sequencing platforms, we prioritized the inclusion of datasets generated by the 10x Genomics platform. Ultimately, we identified 13 target datasets[13–25] to be included in the SexTumorDB and downloaded them for subsequent processing (Table 1). Collectively, these datasets span 13 human tissues and encompass tumors arising in multiple non-reproductive systems, from the lymphatic system (acute myeloid leukemia, follicular lymphoma), digestive system (head and neck squamous cell carcinoma, pancreatic ductal adenocarcinoma, esophageal squamous cell carcinoma, gastric cancer, colorectal cancer, liver cancer), respiratory system (non-small cell lung cancer), endocrine system (papillary thyroid carcinoma), urinary system (bladder cancer, clear cell renal cell carcinoma) to the nervous system (glioblastoma, Fig. 2).

Details of the datasets included in the SexTumorDB are provided in Table 1 and described below. For reading convenience, the "Cohort ID" of relevant datasets (as listed in Table 1) is used to indicate them in the subsequent content.

AML_Galen2019 (acute myeloid leukemia). Galen *et al*. conducted a systematic study on acute myeloid leukemia (AML)[13], a heterogeneous and aggressive blood cancer marked by a high recurrence rate and mortality within 5 years of diagnosis. This research constructed a high-quality dataset of bone marrow aspirates from 16 AML patients through nanowell-based sequencing (Seq-Well). This dataset encompasses a wide range of genetic mutations commonly associated with AML, including DNMT3A, FLT3 and NPM1 mutations. The scRNA-seq dataset generated in this study is publicly available through GSE116256 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116256).

BLCA_Juric2025 (bladder cancer). Bladder cancer (BLCA) typically originates from urothelial or "umbrella" cells. Although it affects both sexes, the incidence in men is three to four times higher than in women. Notably, non-muscle-invasive bladder cancer (NMIBC) accounts for approximately 75–80% of all cases. Juric *et al*. employed single-cell RNA sequencing to profile samples derived from NMIBC patients, comprising 14 treatment-naïve samples[25], providing valuable insights into the heterogeneity and molecular features of NMIBC. The dataset was obtained from the GEO repository GSE269877 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE269877).

ccRCC_Hu2024 (clear cell renal cell carcinoma). Renal cell carcinoma is a common cancer, of which the clear cell renal cell carcinoma (ccRCC) is the most prevalent histologic subtype. Hu *et al*. created a comprehensive snRNA-seq atlas covering 100 treatment-naïve ccRCC samples and 50 matched normal adjacent tissues[14]. This cohort consists of 63 males and 37 females, aligning with the sex distribution in ccRCC. The processed sequencing data from this study are publicly accessible on the Zenodo platform (https://zenodo.org/records/8063124).

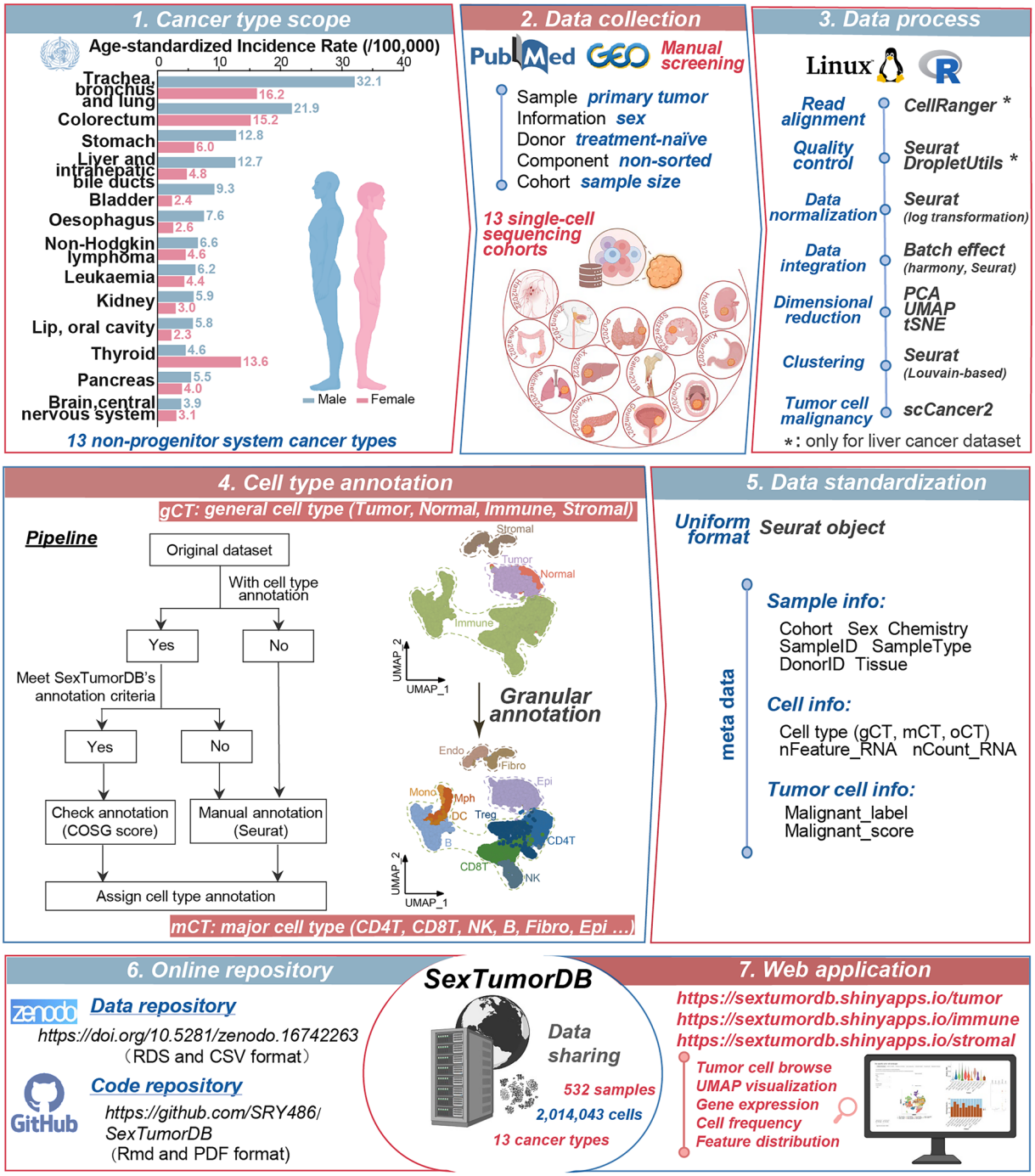

Fig. 1 Schematic overview of the SexTumorDB development workflow. SexTumorDB compiles single-cell and single-nucleus RNA sequencing profiles of 13 cancer types with the top 10 incidence rates in either males or females. The included datasets were curated and processed through the standard pipeline, and were integrated for comprehensive research. The standardized data are shared *via* the Zenodo repository and web applications for easy access and exploration.

CRC_Pelka2021 (colorectal cancer). Colorectal cancer (CRC) is currently the third leading cause of cancer-related death in both male and female populations. In this study, Pelka *et al.* applied single-cell RNA-seq (scRNA-seq) to depict a comprehensive atlas of colorectal tumors as well as samples adjacent to tumors[15]. This dataset extensively covers two genetically distinct CRC subtypes with variable mutation burden, the mismatch repair-deficient (MMRd) subtype and mismatch repair-proficient (MMRp) subtype. This dataset is publicly available on the GEO platform under the accession number GSE178341 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178341).

Cancer type/ subtype	PMID	Authors	Cohort ID	# of cells	# of samples	# of samples (Female, tumor)	# of cells (Female, tumor)	# of samples (Male, tumor)	# of cells (Male, tumor)	# of samples (Female, normal or nomal ajacent)	# of cells (Female, normal or nomal ajacent)	# of samples (Male, normal or nomal ajacent)	# of cells (Male, normal or nomal ajacent)	Sequencing type	Chemistry	#License/ Data Use Agreement	DOI	Source of original data
Acute myeloid leukemia	30827681	Galen *et al.*	AML_ Galen2019	20194	18	5	4877	9	10644	0	0	4	4673	scRNA-seq	Seq-Well	publicly available	https://doi.org/10.1016/j.cell.2019.01.031	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116256
Bladder cancer	41006678	Juric *et al.*	BLCA_ Juric2025	121356	14	4	29501	10	91855	0	0	0	0	scRNA-seq	10x	publicly available	https://doi.org/10.1038/s41698-025-01093-3	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE269877
Clear cell renal cell carcinoma	38361033	Hu *et al.*	ccRCC_ Hu2024	96141	20	6	24263	14	71878	0	0	0	0	snRNA-seq	10x	publicly available	https://doi.org/10.1038/s41588-024-01662-5	https://zenodo.org/record/8063124
Colorectal cancer	34450029	Pelka *et al.*	CRC_ Pelka2021	306203	99	32	98302	32	110020	17	56081	18	41800	scRNA-seq	10x	publicly available	https://doi.org/10.1016/j.cell.2021.08.003	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178341
Esophageal squamous cell carcinoma	34489433	Zhang *et al.*	ESCA_ Zhang2021	198387	64	16	45566	44	136407	0	0	4	16414	scRNA-seq	10x	publicly available	https://doi.org/10.1038/s41467-021-25539-x	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160269
Follicular lymphoma	35687817	Han *et al.*	FL_ Han2022	65946	12	5	31765	4	19736	2	9402	1	5043	scRNA-seq	10x	publicly available	https://doi.org/10.1158/2643-3230.BCD-21-0075	https://cellxgene.cziscience.com/collections/968834a0–1895–40df-8720–666029b3bbac

Continued

Cancer type/ subtype	PMID	Authors	Cohort ID	# of cells	# of samples	# of samples (Female, tumor)	# of cells (Female, tumor)	# of samples (Male, tumor)	# of cells (Male, tumor)	# of samples (Female, normal or nomal ajacent)	# of cells (Female, normal or nomal ajacent)	# of samples (Male, normal or nomal ajacent)	# of cells (Male, normal or nomal ajacent)	Sequencing type	Chemistry	#License/ Data Use Agreement	DOI	Source of original data
Glioblastoma	40346362	Spitzer *et al.*	GBM_ Spitzer2025	70521	20	9	30656	11	39865	0	0	0	0	snRNA-seq	10x	publicly available	https://doi.org/10.1038/s41588-025-02168-4	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE274546
Gastric cancer	34642171	Kumar *et al.*	GC_ Kumar2022	140745	36	9	34399	17	76981	3	15332	7	14033	scRNA-seq	10x	publicly available	https://doi.org/10.1158/2159-8290.CD-21-0683	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183904
Head and neck squamous cell carcinoma	36828832	Choi *et al.*	HNSCC_ Choi2023	8781	10	3	3638	7	5143	0	0	0	0	scRNA-seq	10x	publicly available	https://doi.org/10.1038/s41467-023-36691-x	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE181919
Liver cancer	36352227	Xue *et al.*	LIHC_ Xue2022	223117	26	13	109496	13	113621	0	0	0	0	scRNA-seq	10x	publicly available	https://doi.org/10.1038/s41586-022-05400-x	https://ngdc.cncb.ac.cn/gsa-human/browse/HRA001748
Non-small cell lung cancer	36368318	Salcher *et al.*	NSCLC_ Salcher2022	555200	182	25	71411	49	178949	48	146143	60	158697	scRNA-seq	10x	publicly available	https://doi.org/10.1016/j.ccell.2022.10.008	https://doi.org/10.5281/zenodo.6411867
Pancreatic ductal adenocarcinoma	35902743	Hwang *et al.*	PDAC_ Hwang2022	108964	18	8	51054	10	57910	0	0	0	0	snRNA-seq	10x	publicly available	https://doi.org/10.1038/s41588-022-01134-8	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE202051
Papillary thyroid carcinoma	34663816	Pu *et al.*	THCA_ Pu2021	98488	13	4	26192	3	28806	4	33072	2	10418	scRNA-seq	10x	publicly available	https://doi.org/10.1038/s41467-021-26343-3	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184362

Table 1. Overview of datasets included in the SexTumoDB.

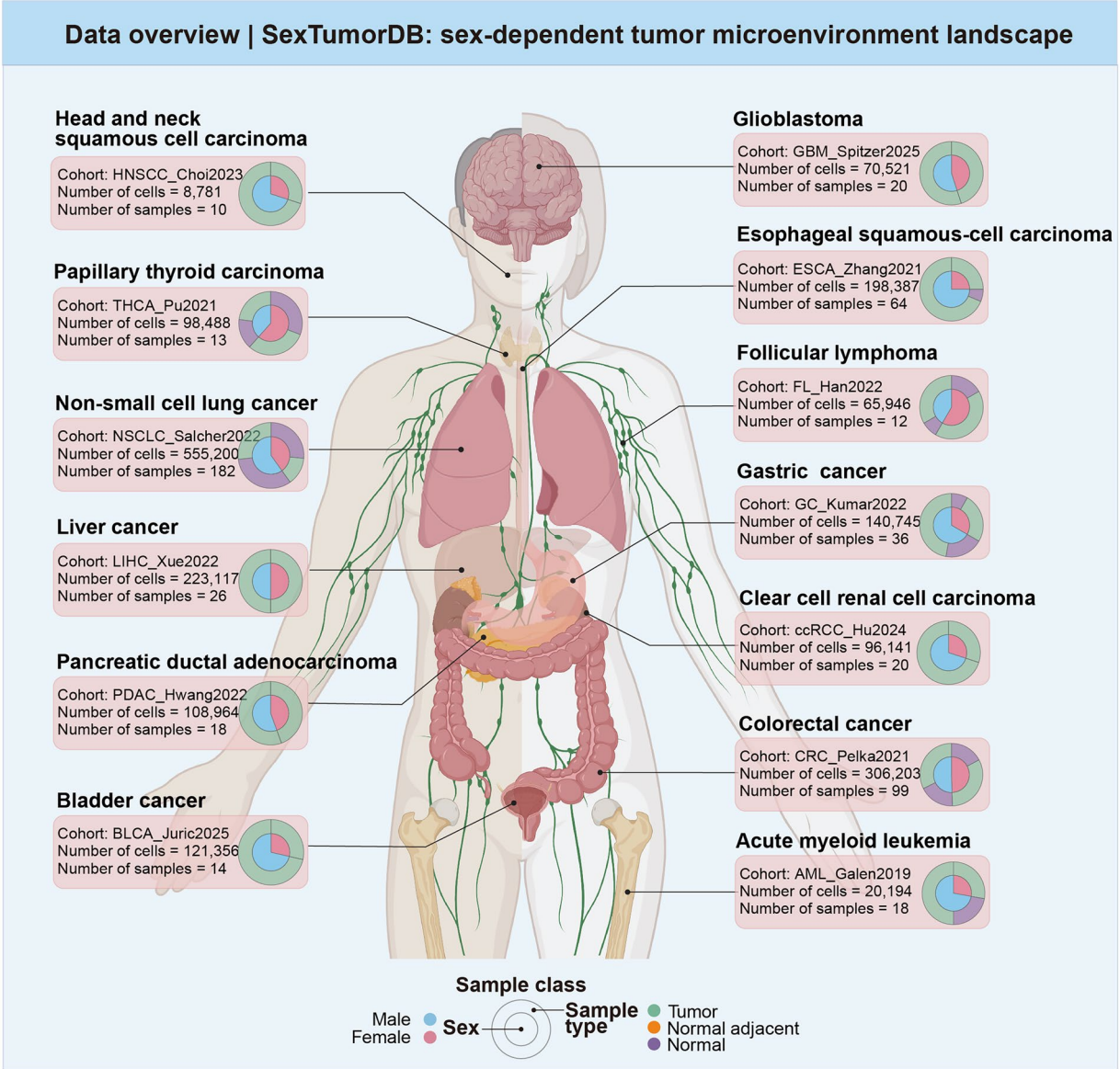

Fig. 2 Overview of cancer cohort datasets compiled in SexTumorDB. SexTumorDB spans 13 human tissues, encompassing tumors arising in multiple non-reproductive systems, including the lymphatic system, respiratory system, digestive system, endocrine system, urinary system, and nervous system. SexTumorDB comprises a total of 532 samples, including 319 male samples and 213 female samples. Among them, 68.05% (362 samples) are tumor samples, including 139 female samples and 223 male samples. This plot illustrates the statistics of cells and samples included in each individual cancer dataset compiled in SexTumorDB.

ESCA_Zhang2021 (esophageal carcinoma). Esophageal squamous-cell carcinoma (ESCC) is the most common histological subtype of esophageal carcinoma (ESCA), comprising approximately 90% of all esophageal cancer cases. In this study, Zhang *et al*. presented a comprehensive single-cell landscape of tumors and adjacent normal samples derived from ESCC patients, furthering our understanding of ESCC[16]. The dataset from this study is publicly available under GSE160269 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160269).

FL_Han2022 (follicular lymphoma). Non-Hodgkin lymphoma (NHL) is a lymphoid tumor that originates from immune cells. The classification of NHL is highly complex, with follicular lymphoma (FL) being the most common indolent NHL subtype. Utilizing the 10x Chromium platform, Han *et al*. constructed a scRNA-seq landscape of 20 FL lymph node biopsies and three reactive lymph nodes[17]. The FL dataset is publicly available in the CELLxGENE database (https://cellxgene.cziscience.com/collections/968834a0-1895-40df-8720-666029b3bbac).

GBM_Spitzer2025 (glioblastoma). Glioblastoma (GBM) is the most prevalent brain tumor, accounting for more than half of all primary malignant tumors in the central nervous system. IDH-wildtype GBM is the most aggressive GBM subtype. In this study, Spitzer *et al*. employed snRNA-seq to analyze 59 paired primary and

recurrent IDH-wildtype GBM samples, providing a high-quality landscape of the GBM ecosystem[18]. The GBM dataset was acquired from GSE274546 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE274546) on the GEO platform.

GC_Kumar2022 (gastric cancer). Gastric cancer (GC) is the fifth leading cause of cancer-related death. Kumar *et al*. generated a large-scale single-cell atlas of GC, encompassing 48 samples from 31 patients across various clinical stages and histologic subtypes[19]. This high-resolution resource enables researchers to uncover additional disease characteristics. The GC dataset is publicly available under accession number GSE183904 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183904).

HNSCC_Choi2023 (head and neck squamous cell carcinoma). Head and neck squamous cell carcinoma (HNSCC) is the most common subtype of head and neck cancer, including oral, pharyngeal and laryngeal cancers. In this study, Choi *et al*. developed a dynamic scRNA-seq profile tracing the evolution from normal tissues to metastatic tumors in 23 HNSCC patients[20]. This dataset was obtained from GSE181919 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE181919). Only samples taken from the oral cavity are included in the SexTumorDB, according to the data collection scope.

LIHC_Xue2022 (liver hepatocellular carcinoma). Liver cancer is a significant global health concern. Primary liver cancer includes three major histological subtypes, hepatocellular carcinoma, intrahepatic cholangiocarcinoma and the combined type. Among these, liver hepatocellular carcinoma (LIHC) is the most common subtype. Xue *et al*. constructed an extensive scRNA-seq atlas of liver cancer, encompassing 160 samples from 124 treatment-naïve patients across all these three subtypes[21]. The raw sequence data generated in this study are deposited in the Genome Sequence Archive for Human (GSA-human) platform under HRA001748 (https://ngdc.cncb.ac.cn/gsa-human/browse/HRA001748). As this is a controlled-access data resource, we obtained access to the dataset upon successful application to the Data Access Committee (DAC) appointed by the data submitter and downloaded the data within the authorized time following DAC approval. To ensure consistency across samples, we performed age and sex matching using the MatchIt R package (v4.7.1) on the LIHC patient-derived samples in the dataset, ultimately selecting 13 male samples and 13 female samples for inclusion in SexTumorDB.

NSCLC_Salcher2022 (non-small cell lung cancer). Lung cancer is the leading cause of cancer-related death around the world. It consists of two subtypes, non-small cell lung cancer (NSCLC) and small cell lung cancer, with NSCLC accounting for the majority of diagnoses. By integrating 556 NSCLC samples from 318 patients spanning 29 datasets, Salcher *et al*. established a large-scale NSCLC atlas[22]. This integrated dataset is publicly available in the Zenodo repository (https://doi.org/10.5281/zenodo.6411867).

PDAC_Hwang2022 (pancreatic ductal adenocarcinoma). Among common solid tumors, pancreatic cancer demonstrates the worst prognosis. Pancreatic ductal adenocarcinoma (PDAC) is the most prevalent subtype of pancreatic cancer. In this study, Hwang *et al*. applied snRNA-seq to develop the molecular stratification of PDAC patients[23]. By enrolling 43 patients with primary PDAC who received neoadjuvant therapy or were treatment-naïve, Hwang *et al*. generated a comprehensive PDAC landscape and characterized the inter- and intra-tumoral heterogeneity. This dataset is publicly available on the GEO platform under GSE202051 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE202051).

THCA_Pu2021 (papillary thyroid carcinoma). Thyroid cancer (THCA) exhibits a distinct propensity for the female population among common solid cancers. Over 80% of thyroid cancers are papillary thyroid carcinoma (PTC). This study presents a scRNA-seq dataset covering the full spectrum of clinical stages in PTC. This data resource is publicly available under accession number GSE184362 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184362).

Data processing. For the sequencing profile of each individual cancer type, we first performed the data processing using the methods outlined below. Of note, certain datasets have already undergone partial processing steps such as read alignment, quality control, the methods described below are optional and may vary depending on the specific dataset. Detailed information on data processing can be found in the GitHub repository (https://github.com/SRY486/SexTumorDB).

Read alignment and quality control. For the dataset providing only raw FASTQ files ("LIHC_Xue2022"), read alignment was performed using the Cell Ranger software (v7.1.0, 10x Genomics) against the pre-built GRCh38 human reference transcriptome (refdata-gex-GRCh38-2020-A, 10x Genomics). The resulting gene-cell matrices were then imported into R (v4.2.3)[27] and processed with the Seurat package (v4.3.0)[28]. To ensure the data consistency, for datasets that utilize Ensembl IDs ("FL_Han2022", "NSCLC_Salcher2022"), we converted them to corresponding official gene symbols. For gene symbols mapping to multiple Ensembl IDs, we summed the corresponding rows in the gene-cell matrices.

Quality control of the "LIHC_Xue2022" dataset was subsequently conducted according to the R scripts provided in the original article (https://github.com/meta-cancer/scPLC). In brief, potential empty droplets and doublets were evaluated by the DropletUtils (v1.18.1)[29] and filtered out. Also, low-quality cells were discarded based on the following criteria: (1) read count (nCount_RNA) $> 30{,}000$; (2) detected genes (nGene) < 500 or $> 6{,}000$; (3) percentage of mitochondrial genes in the total reads (percent.mt) ≥ 50. All other datasets had already undergone read alignment and quality control, they were directly imported into R *via* Seurat package for downstream processing.

Cancer type/subtype	Cohort ID	# of discarded samples (data collection: cellline, treated, sorted, non-10x Genomics, non-primary, samples)	# of discarded cells (data process)	# of discarded cells (cell annotation)	# of discarded cells (unified cell QC)	# of discarded samples (unified sample QC)	# of finally included samples	# of finally included cells
Acute myeloid leukemia	AML_Galen2019	14	0	0	7	2	20194	18
Bladder cancer	BLCA_Juric2025	13	0	9766	5321	0	121356	14
Clear cell renal cell carcinoma	ccRCC_Hu2024	0	0	0	1837	0	96141	20
Colorectal cancer	CRC_Pelka2021	0	0	0	63867	1	306203	99
Esophageal squamous cell carcinoma	ESCA_Zhang2021	0	0	0	10272	0	198387	64
Follicular lymphoma	FL_Han2022	11	0	0	0	0	65946	12
Glioblastoma	GBM_Spitzer2025	101	0	50751	829	0	70521	20
Gastric cancer	GC_Kumar2022	4	0	6042	4152	0	140745	36
Head and neck squamous cell carcinoma	HNSCC_Choi2023	27	0	252	1364	0	8781	10
Liver cancer	LIHC_Xue2022	67	98766	3	791	0	223117	26
Non-small cell lung cancer	NSCLC_Salcher2022	266	0	8452	566	57	555200	182
Pancreatic ductal adenocarcinoma	PDAC_Hwang2022	25	0	0	0	0	108964	18
Papillary thyroid carcinoma	THCA_Pu2021	0	0	0	0	0	98488	13

Table 2. Statistics of samples and cells dicarded from each individual original dataset.

Validation of sex classification. To validate sex classification of the included samples, we first identified genes stably demonstrate a sex-biased expression pattern across 29 human tissues through a published research, which calculated genes differentially expressed by sex in each individual human tissue and provided the fold change (FC) of gene expression in males *versus* females[30]. Genes exhibiting the lowest 10 FC in all 29 tissues is considered as a stable female-biased signature (*XIST*), genes with a top 10 FC in all tissues are identified as a stable male-biased signature (*DDX3Y*, *KDM5D*, *RPS4Y1*, *TXLNG2P*, *USP9Y*, *UTY*, *ZFY*). Also, we obtained the list of genes on chromosome Y (chrY genes) and applied it as another male-biased marker.

Sample filtering. To ensure data quality and facilitate the analysis of relevant data, these datasets were further filtered to retain only samples without pre-treatment from the primary sites of corresponding tumor. Cell-line samples or sorted samples ("AML_Galen2019") were discarded accordingly. Samples derived from non-primary sites were removed ("GC_Kumar2022", "GBM_Spitzer2025", "HNSCC_Choi2023"). Samples with a treatment history were also excluded ("AML_Galen2019", "FL_Han2022", "GBM_Spitzer2025", "PDAC_Hwang2022", "BLCA_Juric2025"). Especially, when processing the "NSCLC_Salcher2022" dataset, we filtered out samples that were not sequenced using the 10x Genomics platform to minimize technical errors across samples (Table 2).

Samples from non-tumor patients with chronic diseases, such as those with chronic obstructive pulmonary disease in the "NSCLC_Salcher2022" dataset, were filtered out. On the other hand, considering that some cancer studies focus on the progressive transition from normal tissue to malignant tumor, non-tumor samples from healthy donors ("normal") or from tumor-adjacent sites ("normal_adjacent") were also retained and are distinguished through the "SampleType" item in the metadata of Seurat object.

Cell integration, visualization and annotation. Data integration and clustering were carried out following the standard pipeline of Seurat for each individual cancer dataset. Briefly, the Seurat objects of samples were first merged into a single integrated Seurat object. Data normalization, highly variable gene (HVG) identification, data scaling were then conducted with the NormalizeData (normalization.method = "LogNormalize", scale.factor = 10000), FindVariableFeatures (selection.method = "vst", nfeatures = 1000~3000), ScaleData (regressing out "nCount_RNA") function. The number of HVGs was determined by dataset size, complexity, and biological heterogeneity, with smaller or less heterogeneous datasets utilizing less HVGs (1000 or 1500) whereas larger or complex datasets using more HVGs (2000 or 3000). Considering the cross-sample differences, batch effects were corrected primarily using the harmony (v0.1.1) toolkit[31], applied to the top 50 principal components from the principal component analysis (PCA) reduction result. For datasets exhibiting extremely high inter-sample variation, the "anchor-based" integration workflow of Seurat was taken. In brief, samples were first processed with data normalization, HVG screening, data scaling and PCA reduction in sequence. A cross-sample anchor set was then identified and used to combine the samples into one integrated Seurat object. Visualization of cells in 2D space was achieved through Uniform Manifold Approximation and Projection (UMAP) dimensional reduction based on the top principal components (PCs). The optimal number of PCs was determined by the smaller value of: (1)

the first PC where cumulative variance exceeded 90% while individual PC variance dropped below 5%; (2) the first PC where the difference in variance between consecutive PCs fell below 0.1.

To enable systematic analyses of sex disparities in the TME, a unified hierarchical cell annotation framework was applied across all SexTumorDB datasets. Three annotation levels were provided, ranging from general to detailed: (1) general cell type (gCT), which categorizes cells into "Tumor", "Normal", "Immune" or "Stromal", summarizing the four main cellular components of the TME[32]; (2) major cell type (mCT), where cells are annotated as epithelial, B, macrophage, fibroblast, etc. Specifically for T cells, given their high abundancy in the TME as well as their crucial role in tumor immunity and immune checkpoint (ICB) therapy[33], we offer a more granular annotation, classifying them into CD4T, CD8T, regulatory T (Treg), etc; (3) original cell type (oCT), which is assigned as the cell type annotation provided by the original data source or manually annotated in this study. It is important to note that, for pre-malignant cells derived from normal or adjacent normal tissues, such as normal epithelial cells from healthy donors in the "NSCLC_Salcher2022" dataset, we classify them under the "Normal" gCT.

Cell annotation was performed using a stepwise strategy integrating prior knowledge and data-driven refinement (Fig. 1). First, marker gene sets were curated from the CellMarker database (http://117.50.127.228/CellMarker/)[34] and original publications (Supplementary Table 1). For datasets with existing annotations, original annotations were assigned as oCT and validated using COSG (v0.9.0)[35] by comparing the top 50 cell type–specific marker genes with the curated markers. When original annotations were ambiguous or incompatible with the unified standard (e.g, "NK/T" cells in the HNSCC dataset), corresponding cells were subsetted for further manual annotation, together with datasets without annotations. The manual annotation was performed using the Seurat pipeline. Briefly, cells were clustered *via* Louvain algorithm-based clustering (FindNeighbors, FindClusters functions), cluster-specific signature genes were then identified with FindAllMarkers (test.use = "wilcox"), cell types were subsequently assigned by matching these signatures to prebuilt marker gene sets. The oCTs passed manual verification or were obtained through manual annotation were then mapped to corresponding mCTs and further categorized under gCTs.

Collectively, we provide a standardized as well as hierarchical cell annotation system, ultimately encompassing 33 mCTs and 237 oCTs under 4 gCTs, offering an extensive view of the cellular diversity across various tumor microenvironments.

Metadata standardization. For the convenience of the scientific community to quickly reuse and analyze the datasets in SexTumorDB, we standardized the format of the datasets before depositing them into online repositories. For each processed dataset, we provide a standard Seurat object with unified metadata, which includes a total of 12 items as described below:

- Sex: sex classification of the donor corresponding to the cell, including "F" (female) and "M" (male).
- Cohort: cancer cohort ID of the data, composed of the cancer type abbreviation, the author name and the publication year of the corresponding research, "THCA_Pu2021" for example.
- Tissue: tissue source of the sample.
- SampleType: classification of the sample type, including "tumor", "normal" and "normal_adjacent", depending on the sampling site and health status of the sample donor.
- SampleID: a unique ID for the sequenced sample.
- DonorID: a unique ID for the sample donor, if not specifically provided in the original research, the SampleID was used as a substitute.
- Chemistry: sequencing platform of the corresponding sample, including Seq-Well (only for "AML_Galen2019") and 10x Genomics (all the other datasets).
- gCT: general cell type, including "Tumor", "Normal", "Immune" and "Stromal".
- mCT: major cell type, a more granular cell type annotation than gCT, classifying cells into a total of 33 cell types, such as "Epi" (epithelial), "Fibro" (fibroblast), "Treg" (regulatory T), "CD8T" ($CD8^+$ T).
- oCT: original cell type, the cell type annotation provided by the original data source or manually annotated in this study.
- nCount_RNA: number of reads detected in the cell.
- nFeature_RNA: number of genes detected in the cell.

Data integration. After preprocessing of each individual cancer dataset, we integrated the quality control metrics (nFeature_RNA, nCount_RNA, percent.mt) across the 13 datasets and performed a comprehensive QC to ensure data consistency. Thresholds for nFeature_RNA and nCount_RNA were defined as three times the interquartile range (IQR) above the third quartile ($Q3 + 3 \times IQR$). As a result, cells with nFeature_RNA $\geq$ 7,938 or nCount_RNA $\geq$ 29,008 and also cells with percent.mt $>$ 40% were uniformly discarded across all the datasets. After that, samples with $<$ 100 cells remaining were further discarded (Table 2).

For a comprehensive overview of sex-dependent tumor microenvironment variations across pan-cancer cohorts at single-cell resolution, we extracted all 362 tumor samples from the 13 datasets and split them according to the gCT annotation. This resulted in three distinct datasets representing the tumor (570,139 cells), immune (757,315 cells), and stromal (175,481 cells) components from all the tumor samples in the SexTumorDB. Only the 13,412 genes detected in all 13 datasets were retained in the integrated datasets.

Given that the large data size might increase the difficulty of data access and analysis, we created a lightweight version of these datasets to facilitate data reuse. For the tumor dataset, samples with more than 1,000 cells were randomly downsampled to 1,000 cells. For the immune and stromal datasets, cells were first grouped by the

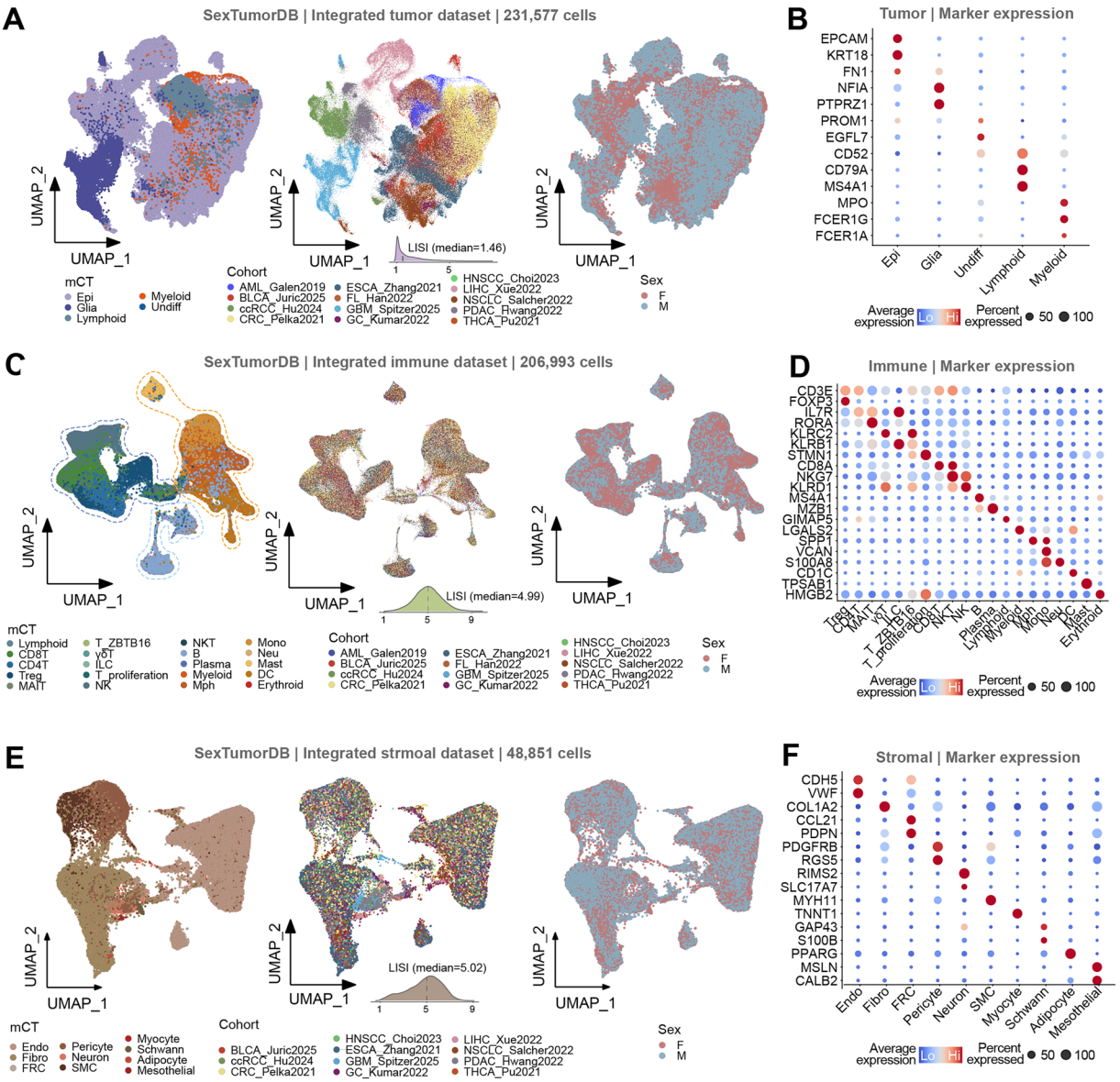

Fig. 3 Lightweight integrated dataset of SexTumorDB. (**A,C,E**) UMAP embeddings of cells in the lightweight integrated dataset for tumor (**A**), immune (**C**) and stromal (**E**) components, which are downsampled from all the tumor samples included in the SexTumorDB. Each dot represents one cell, colored by corresponding major cell type (mCT, left), cohort (middle) or sex classification (right). The bottom ridgeplot displays the LISI score of batch (cohort) distribution, with corresponding median score labeled. UMAP, Uniform manifold approximation and projection; LISI, Local Inverse Simpson's Index; F, female; M, male; Epi, epithelial; Undiff, undifferentiated lineage; Endo, endothelial; Fibro, fibroblast; FRC, fibroblastic reticular cell; SMC, smooth muscle cell; CD8T, $CD8^+$ T; CD4T, $CD4^+$ T; Treg, regulatory T; ILC, innate lymphoid cell; MAIT, mucosal-associated invariant T; Mph, macrophage; Mono, monocyte; DC, dendritic cell; Neu, neutrophil. (**B,D,F**) Bubble heatmap showing the expression pattern of canonical signature genes in manually annotated major cell types in the lightweight integrated dataset for tumor (**B**), immune (**D**) and stromal (**F**) components that downsampled from all the tumor samples included in the SexTumorDB.

combination of their corresponding mCT and SampleID, then groups with more than 100 cells were randomly downsampled to 100 cells. Data integration was subsequently conducted for each individual lightweight dataset. For the immune and stromal datasets, the "anchor-based" integration workflow of Seurat was performed as described above, with data integrated based on identified cross-cohort anchors (FindIntegrationAnchors, IntegrateData function). While for the tumor cell dataset, the harmony tutorial was used, RunHarmony (group.by.vars = " Cohort ") was run to correct batch effect across cohorts. UMAP visualization was subsequently performed on the optimal number of PCs. Furthermore, the LISI (Local Inverse Simpson's Index) score was calculated with compute_lisi function (lisi, v1.0)[31] to quantify the batch effect correction effectiveness based on batch distribution on UMAP embeddings (Fig. 3A,C,E). As a result, we generated three downsampled datasets with

	Location	Content	File format	File name	Brief description
Data repository	https://doi.org/10.5281/zenodo.16742263	Dataset of each individual cancer cohort	Seurat object in RDS	obj.AML_Galen2019.rds	The processed dataset corresponding to each individual cancer cohort included in SexTumorDB
				obj.BLCA_Juric2025.rds	
				obj.ccRCC_Hu2024.rds	
				obj.CRC_Pelka2021.rds	
				obj.ESCA_Zhang2021.rds	
				obj.FL_Han2022.rds	
				obj.GBM_Spitzer2025.rds	
				obj.GC_Kumar2022.rds	
				obj.HNSCC_Choi2023.rds	
				obj.LIHC_Xue2022.rds	
				obj.NSCLC_Salcher2022.rds	
				obj.PDAC_Hwang2022.rds	
				obj.THCA_Pu2021.rds	
		Integrated dataset	Seurat object in RDS	obj.TumorCell.all.rds	The integrated dataset of all the tumor cells included in the SexTumorDB
				obj.TumorCell.diet.rds	The downsampled lightweight datasets of tumor, immune, stromal components from all the tumor samples included in SexTumorDB
				obj.ImmuneCell.diet.rds	
				obj.StromalCell.diet.rds	
		Metadata	Table in RDS	metadata_cell.rds	Metadata of all the cells, samples and cancer cohorts included in SexTumorDB
				metadata_sample.rds	
				metadata_cohort.rds	
			Table in csv	metadata_cell.csv	
				metadata_sample.csv	
				metadata_cohort.csv	
Web application	https://sextumordb.shinyapps.io/tumor/	Integrated tumor cell dataset	APP		Rshiy application for the lightweight tumor cell dataset included in SexTumorDB
	https://sextumordb.shinyapps.io/immune/	Integrated immune cell dataset			Rshiy application for the lightweight immune cell dataset included in SexTumorDB
	https://sextumordb.shinyapps.io/stromal/	Integrated stromal cell dataset			Rshiy application for the lightweight stromal cell dataset included in SexTumorDB
Code repository	https://github.com/SRY486/SexTumorDB	Programmig scripts	Scripts in Rmd and PDF		Programming scripts responsible for data processing, standardization and Rshiny application development precedures

Table 3. Information and details of data and code repository.

231,577 (tumor), 206,993 (immune), 48,851 (stromal) cells each. The lightweight tumor cell dataset was subsequently used to develop the SexTumorDB web application. However, we have also retained the integrated dataset containing all tumor cells and made it publicly available in the Zenodo repository.

Identification of malignant cells. To identify malignant tumor cells in the tumor samples, a lightweight machine-learning model developed in the scCancer2 (v2.3.0) toolkit[36] was utilized. The malignant cell identification module was run on each dataset separately, generating a malignant label that classifies each tumor cell as either "Malignant" or "nonMalignant" along with a scaled malignant score, where a higher score indicates a potentially higher malignancy.

Data Records

The SexTumorDB resource is freely available for download in RDS and CSV formats on the Zenodo platform (https://doi.org/10.5281/zenodo.16742263)[26], containing four types of data to extensively support relevant research: (1) processed and standardized dataset of each individual cancer type (Seurat object in RDS), with unified metadata; (2) an integrated dataset of all the tumor cells from the included tumor samples (Seurat object in RDS), with malignant labels as well as malignant scores of the tumor cells in addition to the standard metadata; (3) downsampled lightweight datasets of the tumor, immune, or stromal components from all the included tumor samples (Seurat object in RDS); (4) metadata for all the cells and samples collected in SexTumorDB, along with summarized statistics for the included cancer cohorts (table in CSV and RDS).

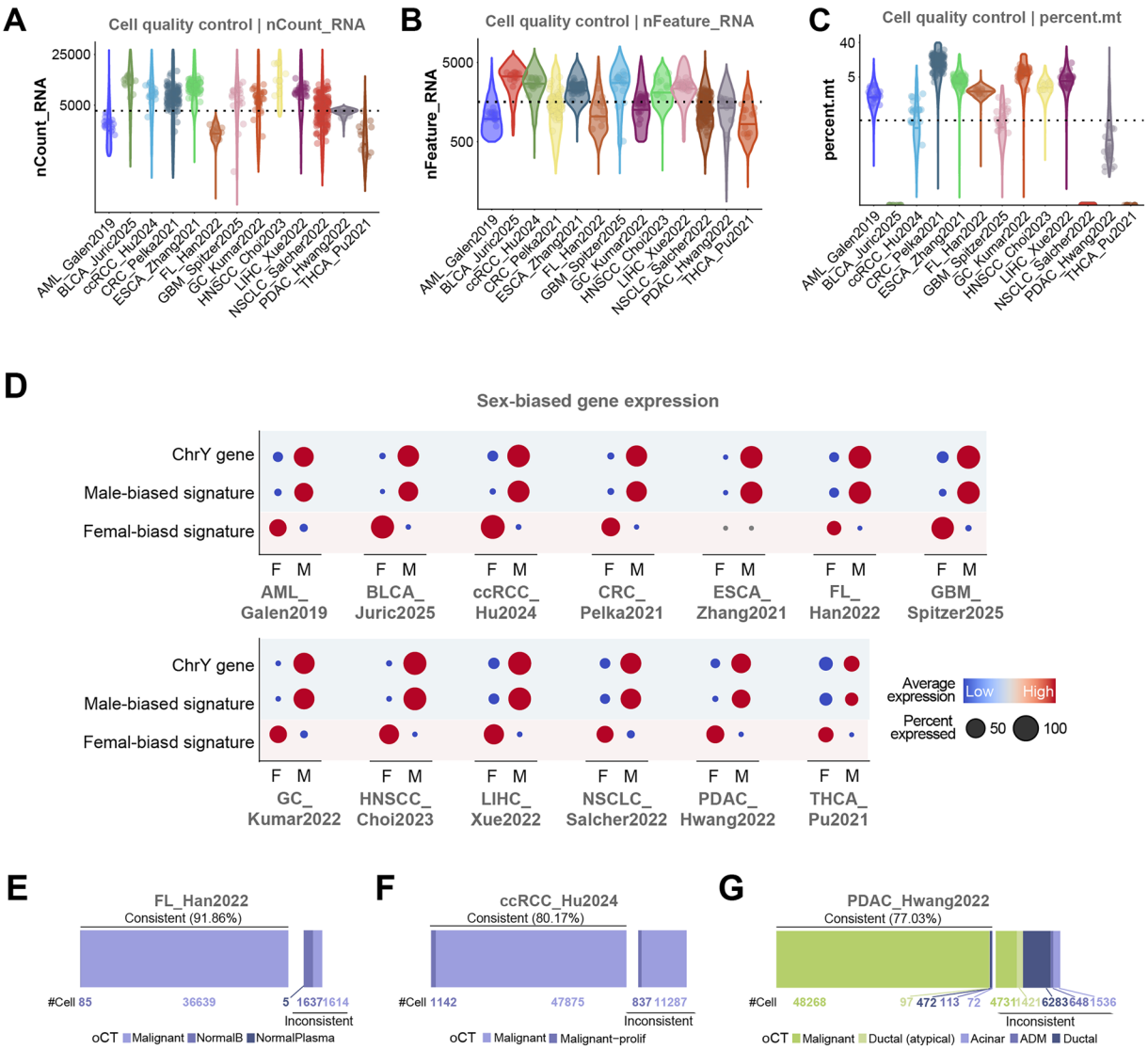

Fig. 4 Technical validation of SexTumorDB data. (**A**–**C**) Violin plots illustrating the distribution of single-cell quality control measures in each individual cancer dataset compiled in SexTumorDB, including nCount_RNA (**A**), nFeature_RNA (**B**) and percent.mt (**C**). Dashed lines indicate the median value of corresponding measure in all cells. Jitter dots represent the average value of the corresponding measure in each individual sample. The Y-axes are log-transformed. (**D**) Bubble plot illustrating the sex-dependent expression patterns of chrY gene, male-biased signature (*DDX3Y*, *KDM5D*, *RPS4Y1*, *TXLNG2P*, *USP9Y*, *UTY*, *ZFY*) and female-biased signature (*XIST*) in female ("F") and male ("M") samples included in SexTumorDB. (**E**–**G**) Bar plot presenting the consistency between scCancer2-predicted malignant label and original malignant label in follicular lymphoma (**E**), clear cell renal cell carcinoma (**F**), and pancreatic ductal adenocarcinoma (**G**) datasets, with the bar color demonstrating the original cell type (oCT) and corresponding cell number labeled below.

Shiny applications are provided for quick data browse and analysis of sex-dependent variations across tumor, immune, and stromal compartments in the SexTumorDB. All the scripts used to generate these datasets can be found in the GitHub repository (https://github.com/SRY486/SexTumorDB), in both Rmd and PDF formats.

More specific details about access to the datasets and web application are provided in Table 3.

Technical Validation

The cancer cohort datasets included in the SexTumorDB were uniformly collected under strict criteria, rigorously curated through manual literature screening, and obtained from high-quality peer-reviewed research. All of the included datasets have undergone meticulous pre-processing or provide detailed scripts for the quality control procedure. To further ensure the data quality and verify the reliability of the SexTumorDB data resource, we performed a series of technical validation procedures using various measures throughout the process of data collection and data processing.

To address the data quality of SexTumorDB, we first checked the sequencing quality of datasets corresponding to each individual cancer cohort at single-cell level, using classic scRNA-seq quality control parameters,

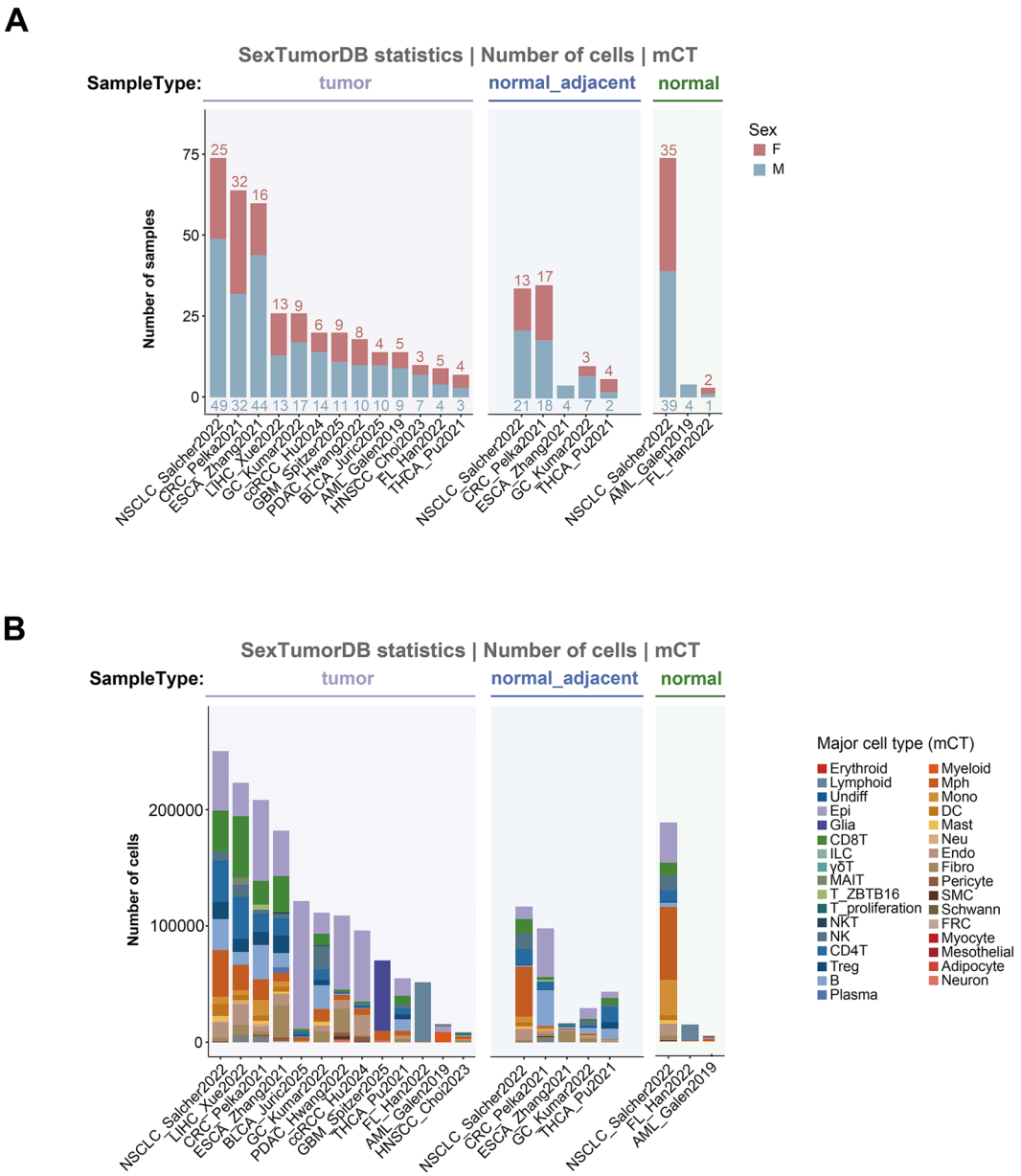

Fig. 5 Statistics of the SexTumorDB datasets. (**A**) Stacked bar plots illustrating the distribution of female ("F", colored in red) and male ("M", colored in blue) samples in the SexTumorDB datasets, grouped by sample type, with corresponding sample number labeled. (**B**) Stacked bar plots presenting the cellular composition in the SexTumorDB datasets, with the bar color demonstrating the manually annotated major cell type (mCT) generated in this study.

including nCount_RNA, nFeature_RNA and percent.mt (Fig. 4A–C). All the included cells exhibit the nCount_RNA range from 214 to 29,007 (median = 4,199), the nFeature_RNA range from 88 to 7,936 (median = 1,589) and the percent.mt range from 0 to 40 (median = 0.37). These metrics collectively support the reliability of sequencing profiles compiled in SexTumorDB.

Given that sex classification of samples is especially important for the reuse of SexTumorDB resource, we evaluated the accuracy of sex classification of included samples *via* the expression pattern of pre-identified sex-biased signatures, including the female-biased signature (*XIST*), male-biased signature (*DDX3Y*, *KDM5D*, *RPS4Y1*, *TXLNG2P*, *USP9Y*, *UTY*, *ZFY*), and the genes on chromosome Y (chrY). All the included datasets demonstrate significant female-specific expression of *XIST* as well as male-specific expression of the male-biased signature and chrY gene (Fig. 4D), validating the accuracy of the sex classification provided in SexTumorDB.

The quality of cell type annotation was verified through the expression pattern of canonical marker genes (Supplementary Table 1) in the integrated datasets, further affirming the reliability of SexTumorDB datasets (Fig. 3B,D,F).

To validate the accuracy of the scCancer2-predicted malignancy labels, we systematically evaluated the consistency between the scCancer2-predicted label and the original malignant label in datasets that include a "Malignant" classification in oCT. The results demonstrated accuracy rates of 91.86% for the "FL_Han2022" dataset, 80.17% for the "ccRCC_Hu2024" dataset, and 77.03% for the "PDAC_Hwang2022" dataset (Fig. 4E–G), confirming the usability of the malignant label provided in SexTumorDB.

Statistics of sample distribution (Fig. 5A) and cell composition (Fig. 5B) across the SexTumorDB datasets demonstrate the comprehensiveness of the data coverage in SexTumorDB. All the datasets include at least three male tumor samples as well as three female tumor samples. These datasets cover 33 major cell types, spanning 13 human tissues from multiple non-reproductive systems. The extensive sample coverage validates that the SexTumorDB enables a comprehensive exploration of sexual dimorphisms across various tumor microenvironments.

Collectively, these findings underscore the reliability and accuracy of SexTumorDB.

Usage Notes

To share this data source with research fellows and to facilitate studies on the sex-dependent disparities across various tumor microenvironments, we uploaded all the standardized SexTumorDB datasets in RDS format to the Zenodo repository (https://doi.org/10.5281/zenodo.16742263). At the same time, considering that some researchers may not have access to high-performance servers, downsampled integrative datasets for the tumor, immune, and stromal components from all the tumor samples included in SexTumorDB can also be accessed in the Zenodo repository. Moreover, given that some researchers may lack experience in bioinformatics or R programming, we developed the SexTumorDB web applications that allow users to explore sex disparities across various tumor microenvironments in a more convenient way. Researchers can interactively explore sex-dependent features in the tumor (https://sextumordb.shinyapps.io/tumor), immune (https://sextumordb.shinyapps.io/immune), and stromal compartments (https://sextumordb.shinyapps.io/stromal) through the website. Users should be aware that certain cancer cohorts included in SexTumorDB exhibit sex imbalance at the sample size, which reflects both real-world epidemiological disparities and limitations of currently available public single-cell datasets, we recommend that users interpret sex-associated findings from such cohorts with caution.

Data availability

All the SexTumorDB datasets are available in the Zenodo repository (https://doi.org/10.5281/zenodo.16742263). Details about access to SexTumorDB datasets are provided in Table 3.

Code availability

The source codes used to generate the SexTumorDB datasets are available in the GitHub repository (https://github.com/SRY486/SexTumorDB), with all the packages and their versions also listed.

Received: 27 August 2025; Accepted: 23 January 2026;
Published online: 25 February 2026

References

1. Zhang, Y. & Li, X. The impact of sex on metastasis. *Nature reviews. Cancer* **24**, 647–648, https://doi.org/10.1038/s41568-024-00725-3 (2024).
2. Swaby, A., Atallah, A., Varol, O., Cristea, A. & Quail, D. F. Lifestyle and host determinants of antitumor immunity and cancer health disparities. *Trends in cancer* **9**, 1019–1040, https://doi.org/10.1016/j.trecan.2023.08.007 (2023).
3. Clocchiatti, A., Cora, E., Zhang, Y. & Dotto, G. P. Sexual dimorphism in cancer. *Nature reviews. Cancer* **16**, 330–339, https://doi.org/10.1038/nrc.2016.30 (2016).
4. Bray, F. *et al*. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **74**, 229–263, https://doi.org/10.3322/caac.21834 (2024).
5. Haupt, S., Caramia, F., Klein, S. L., Rubin, J. B. & Haupt, Y. Sex disparities matter in cancer development and therapy. *Nature reviews. Cancer* **21**, 393–407, https://doi.org/10.1038/s41568-021-00348-y (2021).
6. Castro, A. *et al*. Strength of immune selection in tumors varies with sex and age. *Nature communications* **11**, 4128, https://doi.org/10.1038/s41467-020-17981-0 (2020).
7. Özdemir, B. C. Removing barriers to address sex differences in anticancer drug toxicity. *Nature reviews. Cancer* **24**, 161–162, https://doi.org/10.1038/s41568-023-00651-w (2024).
8. Liu, L., Sun, P. & Zhang, W. A pan-cancer interrogation of intronic polyadenylation and its association with cancer characteristics. *Briefings in bioinformatics* **25**, https://doi.org/10.1093/bib/bbae376 (2024).
9. Han, J. *et al*. Pan-cancer analysis reveals sex-specific signatures in the tumor microenvironment. *Molecular oncology* **16**, 2153–2173, https://doi.org/10.1002/1878-0261.13203 (2022).
10. Yuan, Y. *et al*. Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. *Cancer cell* **29**, 711–722, https://doi.org/10.1016/j.ccell.2016.04.001 (2016).
11. Li, C. H. *et al*. Sex differences in oncogenic mutational processes. *Nature communications* **11**, 4330, https://doi.org/10.1038/s41467-020-17359-2 (2020).
12. Tirosh, I. & Suva, M. L. Cancer cell states: Lessons from ten years of single-cell RNA-sequencing of human tumors. *Cancer cell* **42**, 1497–1506, https://doi.org/10.1016/j.ccell.2024.08.005 (2024).
13. van Galen, P. *et al*. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265–1281.e1224, https://doi.org/10.1016/j.cell.2019.01.031 (2019).
14. Hu, J. *et al*. Multi-omic profiling of clear cell renal cell carcinoma identifies metabolic reprogramming associated with disease progression. *Nature genetics* **56**, 442–457, https://doi.org/10.1038/s41588-024-01662-5 (2024).
15. Pelka, K. *et al*. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734–4752.e4720, https://doi.org/10.1016/j.cell.2021.08.003 (2021).
16. Zhang, X. *et al*. Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis. *Nature communications* **12**, 5291, https://doi.org/10.1038/s41467-021-25539-x (2021).

17. Han, G. *et al*. Follicular Lymphoma Microenvironment Characteristics Associated with Tumor Cell Mutations and MHC Class II Expression. *Blood cancer discovery* **3**, 428–443, https://doi.org/10.1158/2643-3230.Bcd-21-0075 (2022).
18. Spitzer, A. *et al*. Deciphering the longitudinal trajectories of glioblastoma ecosystems by integrative single-cell genomics. *Nature genetics* **57**, 1168–1178, https://doi.org/10.1038/s41588-025-02168-4 (2025).
19. Kumar, V. *et al*. Single-Cell Atlas of Lineage States, Tumor Microenvironment, and Subtype-Specific Expression Programs in Gastric Cancer. *Cancer discovery* **12**, 670–691, https://doi.org/10.1158/2159-8290.Cd-21-0683 (2022).
20. Choi, J. H. *et al*. Single-cell transcriptome profiling of the stepwise progression of head and neck cancer. *Nature communications* **14**, 1055, https://doi.org/10.1038/s41467-023-36691-x (2023).
21. Xue, R. *et al*. Liver tumour immune microenvironment subtypes and neutrophil heterogeneity. *Nature* **612**, 141–147, https://doi.org/10.1038/s41586-022-05400-x (2022).
22. Salcher, S. *et al*. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer cell* **40**, 1503–1520.e1508, https://doi.org/10.1016/j.ccell.2022.10.008 (2022).
23. Hwang, W. L. *et al*. Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment. *Nature genetics* **54**, 1178–1191, https://doi.org/10.1038/s41588-022-01134-8 (2022).
24. Pu, W. *et al*. Single-cell transcriptomic analysis of the tumor ecosystems underlying initiation and progression of papillary thyroid carcinoma. *Nature communications* **12**, 6058, https://doi.org/10.1038/s41467-021-26343-3 (2021).
25. Juric, I. *et al*. Single-cell RNA-sequencing of BCG naïve and recurrent non-muscle invasive bladder cancer reveals a CD6/ALCAM-mediated immune-suppressive pathway. *NPJ precision oncology* **9**, 318, https://doi.org/10.1038/s41698-025-01093-3 (2025).
26. Sun, R. SexTumorDB: a comprehensive resource of sex-dependent tumor landscape at single-cell resolution. *Zenodo* https://doi.org/10.5281/zenodo.16742263 (2025).
27. R Core Team, R. (Citeseer, 2020).
28. Hao, Y. *et al*. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e3529, https://doi.org/10.1016/j.cell.2021.04.048 (2021).
29. Lun, A. T. L. *et al*. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome biology* **20**, 63, https://doi.org/10.1186/s13059-019-1662-y (2019).
30. Lopes-Ramos, C. M. *et al*. Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. *Cell reports* **31**, 107795, https://doi.org/10.1016/j.celrep.2020.107795 (2020).
31. Korsunsky, I. *et al*. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**, 1289–1296, https://doi.org/10.1038/s41592-019-0619-0 (2019).
32. Harris, M. A. *et al*. Towards targeting the breast cancer immune microenvironment. *Nature reviews. Cancer* **24**, 554–577, https://doi.org/10.1038/s41568-024-00714-6 (2024).
33. Elhanani, O., Ben-Uri, R. & Keren, L. Spatial profiling technologies illuminate the tumor microenvironment. *Cancer cell* **41**, 404–420, https://doi.org/10.1016/j.ccell.2023.01.010 (2023).
34. Hu, C. *et al*. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic acids research* **51**, D870–d876, https://doi.org/10.1093/nar/gkac947 (2023).
35. Dai, M., Pei, X. & Wang, X. J. Accurate and fast cell marker gene identification with COSG. *Briefings in bioinformatics* **23**, https://doi.org/10.1093/bib/bbab579 (2022).
36. Chen, Z. *et al*. scCancer2: data-driven in-depth annotations of the tumor microenvironment at single-level resolution. *Bioinformatics (Oxford, England)* **40**, https://doi.org/10.1093/bioinformatics/btae028 (2024).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (82025017 to D.W.), and the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (LHDMD23H160002 to D.W.). This work was also supported by Provincial Key Laboratory of Precise Diagnosis and Treatment of Abdominal Infection, Sir Run Run Shaw Hospital, and Zhejiang University School of Medicine.

Author contributions

R.S. designed the research; R.S. and Q.D. performed data curation and processing of relevant datasets; R.S., Q.D. and D.W. wrote the manuscript; R.S. and D.W. supervised this project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-06707-4.

Correspondence and requests for materials should be addressed to R.S. or D.W.