scientific **data**

Check for updates

OPEN

DATA DESCRIPTOR

A global database for modeling tumor-immune cell communication

Yunjin Xie[1,4], Weiwei Zhou[1,4], Jingyi Shi[1,4], Mengjia Xu[1], Zijing Lin[2], Donghao Li[1], Jianing Li[1], Shujun Cheng[1,3 ✉], Tingting Shao[1 ✉] & Juan Xu [1 ✉]

Communications between tumor cells and surrounding immune cells help shape the tumor immunity continuum. Recent breakthroughs in high-throughput technologies as well as computational algorithms had reported many important tumor-immune cell (TIC) communications, which were scattered in thousands of published studies and impeded systematical characterization of the TIC communications across cancer. Here, a comprehensive database, TICCom, was developed to model TIC communications, containing 739 experimentally-validated or manually-curated interactions collected from more than 3,000 literatures as well as 4,537,709 predicted interactions inferred via six computational algorithms by reanalyzing 32 scRNA-seq datasets and bulk RNA-seq data across 25 cancer types. The communications between tumor cells and 14 types of immune cells were characterized, and the involved ligand-receptor interactions were further integrated. 14190 *human* and 3650 *mouse* integrated ligand-receptor interactions with supplemented corresponding function information were also stored in the TICCom database. Our database would serve as a valuable resource for investigating TIC communications.

Background & Summary

Modulating the patient immune system with immunotherapy had revolutionized cancer therapy, and led to durable remissions across various cancer types[1,2]. The communications within or between the tumors and surrounding immune cells helped shape the tumor immunity continuum through chemokine-receptor signaling, and may contribute to different responses to immunotherapies[3–5]. Recent breakthroughs in cancer immunotherapy and decreasing costs of high-throughput technologies had sparked intensive research into tumor-immune cell (TIC) interactions. Single-cell RNA-seq (scRNA-seq) had been widely used to explore the cell composition of the tumor microenvironment in various cancer types, as well as the communication within or among these compositions[6–8]. Indeed, many studies had reported the important roles of TIC communications; however, this experimentally supported communication information was hidden in thousands of published studies[9–11]. These fragmented and even inconsistent publications were obstacles to characterizing TIC communications in the tumor microenvironment in both pan-cancer and tissue-specific contexts. Notably, no database was developed to collect these latest and experimentally supported TIC associations.

The communication between tumors and immune cells in the local tumor microenvironment started with the binding of a ligand to its receptor and the activation of specific cell signaling pathway[9,10,12]. Thus, collecting ligand-receptor interactions was fundamental to understanding TIC communication. For example, CellPhoneDB provided a resource of ligands, receptors, and their interactions, which took into account the subunit architecture of receptors[13]. In addition, many ligand-receptor interactions were collected by different prediction methods and resources of cell-cell communications, such as iTALK[14], ICELLNET[15], CellChat[16]. Thus, assembling these ligand-receptor interactions was an urgent task and would improve the prediction accuracy of cell communication.

To address this gap, we developed a comprehensive resource called TICCom to collect and integrate TIC communications. TICCom included not only communications supported by experiments and manual curation from the published literature, but also predicted results by several commonly used computational methods

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China. [2]Endocrinology department, the first affiliated hospital of Harbin Medical University, Harbin, 150081, China. [3]State Key Laboratory of Molecular Oncology, Department of Etiology and Carcinogenesis, Cancer Institute and Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, 100021, China. [4]These authors contributed equally: Yunjin Xie, Weiwei Zhou, Jingyi Shi. ✉e-mail: chengshj@263.net.cn; shaotingting@hrbmu.edu.cn; xujuanbiocc@ems.hrbmu.edu.cn
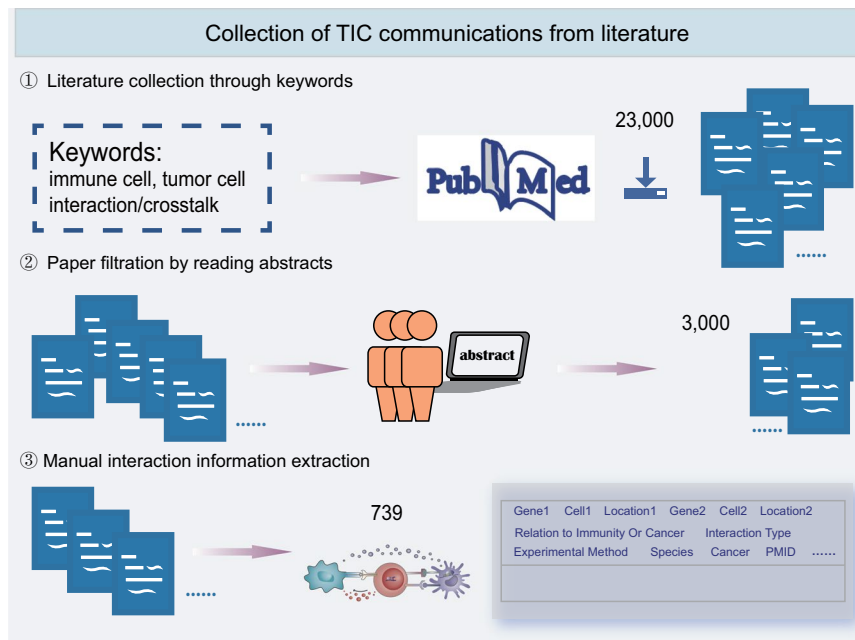
Fig. 1 Collection of TIC communications from literature. Firstly, ~23,000 articles were downloaded from PubMed and retrieved using keywords. Secondly, three researchers carefully read abstracts, and 3000 articles were retained. Finally, detailed information on tumor-immune cell communication was extracted.

based on integrative ligand-receptor interactions and verified TIC interactions. Moreover, the TIC communications were further classified into three types based on the interaction model. We further manually labeled all the ligand-receptor pairs functionally, and much more other information was also provided, including expression of TIC communications across 33 major cancer types, their potential as prognostic markers and TF/miRNA regulation. Several flexible tools were developed to aid retrieval and analysis. TICCom would serve as a valuable resource for investigating communications between tumors and immune cells and greatly extend our understanding of cancer immunotherapy.

Methods

Collection of TIC communications from literature. Firstly, an extensive literature query of the PubMed database was performed using a list of keywords, such as 'immune cell', 'tumor cell', 'crosstalk' and 'interaction'. ~23,000 references were retrieved, the titles and abstracts of which were downloaded. Secondly, we filtered the papers related to tumor-immune interactions by reading the abstracts. Lastly, more than 3,000 literatures remained, and the TIC interaction information was manually extracted from these literatures, including interaction gene pairs, functions, subcellular localizations, experimental methods, descriptions of interactions, titles and PMIDs of literatures, and other details (Fig. 1).

Integration of ligand-receptor interactions data. 14,190 *human* ligand-receptor (LR) pairs were collected from seven studies[7,14–19], while 3,650 *mouse* LR pairs were collected from two studies[19,20]. Genes were represented by Ensembl gene IDs. Considering the immunogenicity of LR interactions, we classified LR pairs into 10 groups: notch signaling, antigen binding, neuropeptide, hormone, growth factor, interferon, interleukin, tumor necrosis factor, chemokine, and cytokine by simultaneously annotating ligands and receptors to relevant GO terms obtained from MSigDB[21]. For example, if a ligand and its coupled receptor were annotated to cytokine-related GO term groups, this pair was assigned to the cytokine group. Unsuccessfully assigned LR pairs were finally appointed to the 'other' group. In order to facilitate the further understanding of the confidence of LR interactions, these integrated interactions were grouped into two subclasses based on their identification methods in previous studies. The interactions were labeled 'manually curated' if they were supported by experimental data or manual annotation from the literature in at least one dataset, and others were grouped into the predicted subclass.

Cancer transcriptome datasets. Bulk RNA-seq data were collected and unified from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/), the International Cancer Genome Consortium (ICGC, https://dcc.icgc.org/)[22], and the EMBL-EBI Expression Atlas (https://www.ebi.ac.uk/gxa/home)[23], including 12,914 samples of 25 cancer types. The cancer transcriptome datasets are shown in Supplementary Table 1. The count matrix of genes was quantified as fragments per kilobase per million reads mapped (FPKM). Genes whose expression was zero in more than 30% of samples were excluded. The expression was log2(FPKM + 0.05) normalized. In addition, 32 single-cell RNA-seq datasets from 13 cancer types having both tumor cells and immune cells were retrieved from both the NCBI Gene Expression Omnibus (GEO)[24,25] and the TISCH database[26]. The scRNA-seq datasets are displayed in Table 1. As previous works[27–29], we predicted cell-cell interactions separately based on

Cancer Type	Cancer Type Detailed	Source	Source2
Bladder Cancer	Bladder Cancer	GSE145137[37]	GEO[24,25]
Leukemia	Acute Lymphocytic Leukemia	GSE132509[38]	TISCH[26]
	Acute Erythroid Leukemia	GSE142213[39]	TISCH
	Acute Myeloid Leukemia	GSE116256[40]	GEO
Breast Cancer	Breast Cancer	GSE143423[41]	TISCH
		GSE75688[42]	GEO
		SRP114962[43]	TISCH
Brain Cancer	Glioma	GSE102130[44]	TISCH
		GSE103224[45]	TISCH
		GSE138794[46]	TISCH
		GSE139448[47]	TISCH
		GSE141982[48]	GEO
		GSE70630[49]	TISCH
		GSE84465[50]	GEO
		GSE89567[51]	TISCH
Head and Neck Cancer	Head and Neck Squamous Cell Carcinoma	GSE103322[52]	GEO
Colorectal Cancer	Colorectal Cancer	GSE146771[53]	GEO
Liver Cancer	Liver Cancer	GSE125449[54]	GEO
Lung Cancer	Lung Adenocarcinoma	GSE131907[55]	GEO
	Non Small Cell Lung Carcinoma	EMTAB6149[56]	TISCH
		GSE117570[57]	TISCH
		GSE127465[58]	TISCH
		GSE143423[41]	TISCH
Neuroendocrine Cancer	Neuroendocrine Cancer	GSE140312[59]	TISCH
Ovarian Cancer	Ovarian Cancer	GSE118828[60]	TISCH
Pancreatic Cancer	Pancreatic Adenocarcinoma	CRA001160[61]	TISCH
		GSE111672[62]	TISCH
Skin Cancer	Skin Cutaneous Melanoma	GSE115978[63]	GEO
		GSE72056[64]	GEO
	Basal Cell Carcinoma	GSE123813[65]	GEO
	Merkel Cell Carcinoma	GSE117988[66]	GEO
	Uveal Melanoma	GSE139829[67]	TISCH
Gastric Cancer	Early Gastric Cancer	GSE134520[68]	GEO

Table 1. scRNA-seq data obtained from GEO and TISCH.

each scRNA-seq dataset containing more than 500 cells or bulk RNA-seq dataset with at least six samples. Cancer categories were carefully unified and listed in Supplementary Table 2.

Computationally predicting TIC communications. TICCom provided inferred cell-cell communication based on 32 scRNA-seq datasets of 13 cancer types through five popular computational methods according to their corresponding standardized processes, including iTALK-top[14], iTALK-DEG[14], CellTalker[30], ICELLNET[15], and NicheNet[17]. The inferred TIC communication was characterized by providing detailed information, including gene names, functions, and labels (evidence) for prediction or manual curation, as well as computational methods and cancer types.

Interaction strengths of TIC communication based on bulk RNA-seq data and the integrated LR interactions were estimated by TItalk provided by TICCom. All genes were ranked in descending order by their expression for each sample. Interaction strength of a pair of interacting genes ISg for a sample n was defined as follows:

$$ISg_n = (rank_{n1} + rank_{n2}) \times (10 - abs(rank_{n1} - rank_{n2}))$$

where $rank_{n1}$ and $rank_{n2}$ were the positions of two interacting genes in the sorted vectors according to their expression in $sample_n$ respectively. The statistical significance was calculated as the probability of observing a lower interaction strength than the true one through 1000 random samplings. The datasets consisted of predicted interaction strengths and p values of gene interactions occurring between tumor and immune cells across cancer types.

miRNA-target and TF-gene interactions. 17,723 miRNA-target interactions were downloaded from starBase[31], and 29,251 TF-gene interactions were downloaded from TRRUST[32], HTRIdb[33], and ORTI[34].
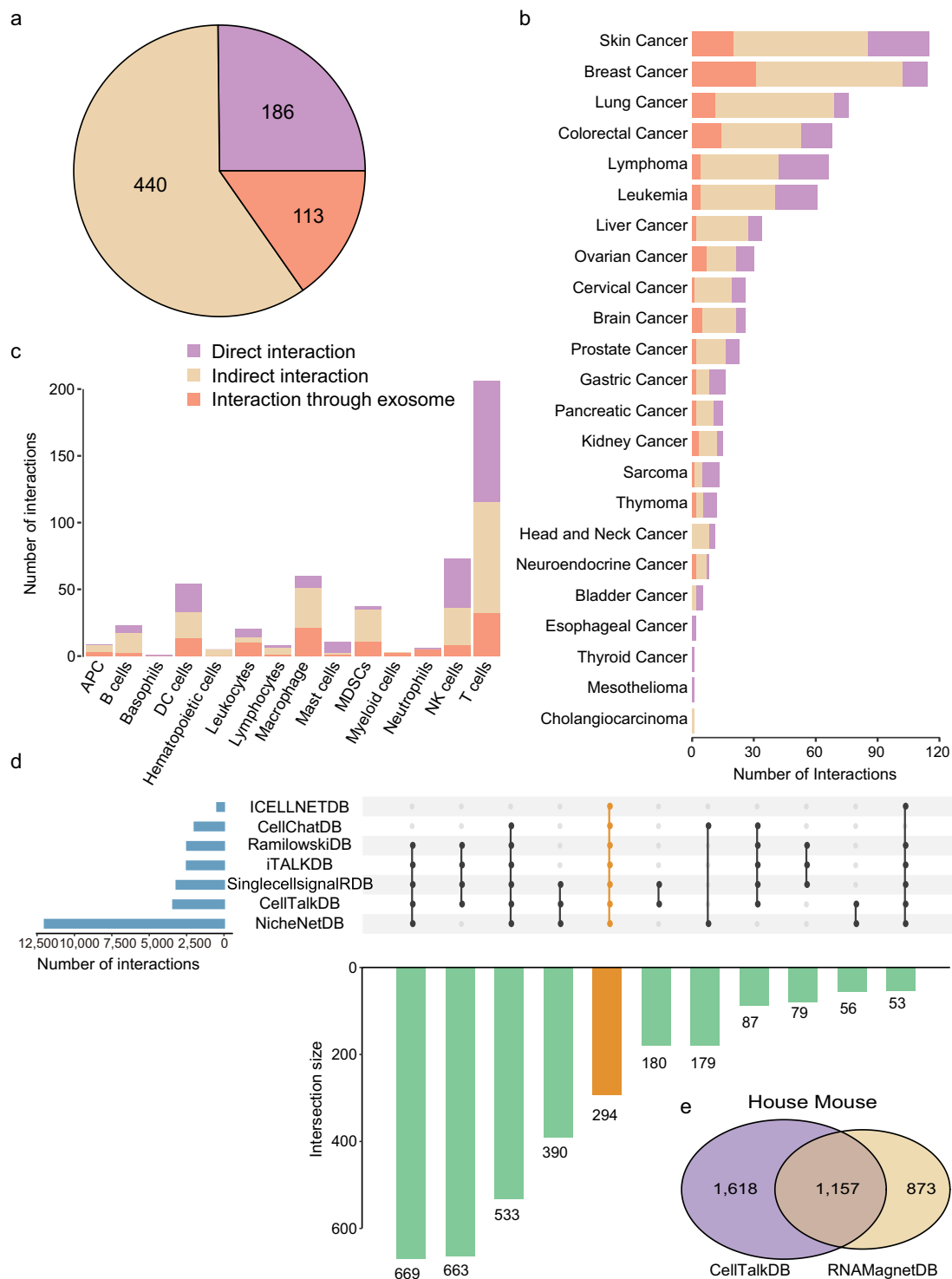
Fig. 2 The resources in TICCom. (**a**) The number of experimentally-verified tumor-immune cell interactions of different interaction types. (**b**) The number of verified tumor-immune cell interactions occurred in 23 cancer types. (**c**) The number of verified tumor-immune cell interactions occurred in 14 immune cells. (**d**) The total number of ligand-receptor interactions across seven *human* datasets was displayed in a bar plot on the left. The seven *human* datasets were represented by dots in the corresponding rows of the dot matrix, which also showed the intersection set of the datasets in its column. The intersection set size was depicted by the bar plot at the bottom. (**e**) The Venn plot showed the number of ligand-receptor interactions shared by the two *mouse* datasets, CellTalkDB and RNAMagnetDB. The left and right circles indicated the number of ligand-receptor interactions in CellTalkDB and RNAMagnetDB, respectively.
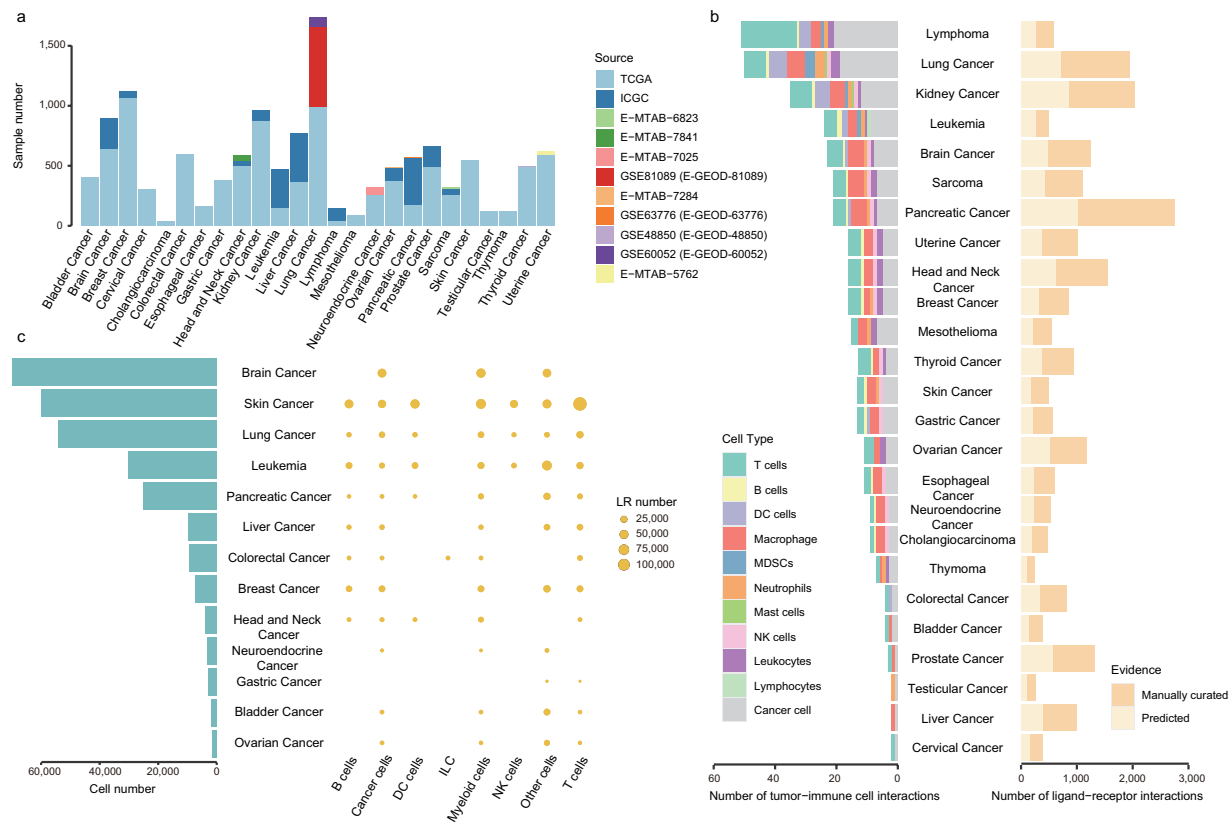
Fig. 3 The predicted interactions stored in TICCom. (**a**) The number of samples in each cancer type from different bulk RNA-seq resources. The datasets with serial numbers containing MTAB or GSE were obtained from the EMBL-EBI Expression Atlas. (**b**) The number of predicted verified tumor-immune cell direct interactions and that of predicted ligand-receptor interactions in bulk RNA-seq data from 25 cancer types. The significant p values of TCGA and ICGC cancer were set at 0.05; however, the significant p value of EMBL was set at 0.1 because of the smaller number of samples. (**c**) The bubble plot showed the number of predicted ligand-receptor interactions that occurred in each cancer and each immune cell inferred by five algorithms based on 32 scRNA-seq datasets from 13 cancer types and the integrated ligand-receptor interactions. The bar plot showed the number of cells in each cancer. The size of the bubble indicated the number of predicted ligand-receptor interactions.

Data Records

These datasets can be obtained from Figshare[35] and from the download page of TICCom (http://bio-bigdata. hrbmu.edu.cn/TICCom/). The R codes used to generate datasets in TICCom were shared on Github (https:// github.com/yunjinxie/TICCom-dataset).

There were five CSV files in the database. All the deposited data was processed, and all the sources were openly available. The CSV table 'Experimentally verified TIC communication' contained detailed information on experimentally-verified TIC communications, including gene symbols, cell types, interaction types, species, experiments, interaction comments, and original reference information. Integrated ligand-receptor interactions were displayed in the CSV table 'Integrated ligand-receptor interactions', including gene symbols, functions, sources of LR pairs, and evidence. Predicted TIC communications based on bulk RNA-seq were stored in the CSV tables 'Predicted TIC communication based on experimentally verified TICs and bulk RNA-seq' and 'Predicted TIC communication based on ligand-receptor interactions and bulk RNA-seq'. Information included gene symbols, cell types, cancer types, interaction strengths, and p values in the former dataset. However, in the latter dataset, ligands, receptors, functions, evidence, cancer types, interaction strengths, and p values were recorded. The CSV table 'Predicted TIC communication based on scRNA-seq' contained detailed information about predicted TIC interactions inferred by five algorithms using 32 scRNA-seq datasets and the integrated ligand-receptor interactions. Ligands, receptors, cell types, prediction methods, evidence, datasets, cancer types, and cancer subtypes were included in this dataset. In Supplementary Table 3, the columns of five CSV files are listed.

Technical Validation

In order to validate the accuracy of experimentally supported TIC communications, the process of data extraction was performed independently by Y.X., J.S., and M.X. and subsequently cross-checked. The resolution of any disagreements regarding data extraction was based on consensus. Information retrieval was done manually. The dataset included 739 experimentally-verified TIC interactions from *human* and *mouse*, covering 26 years
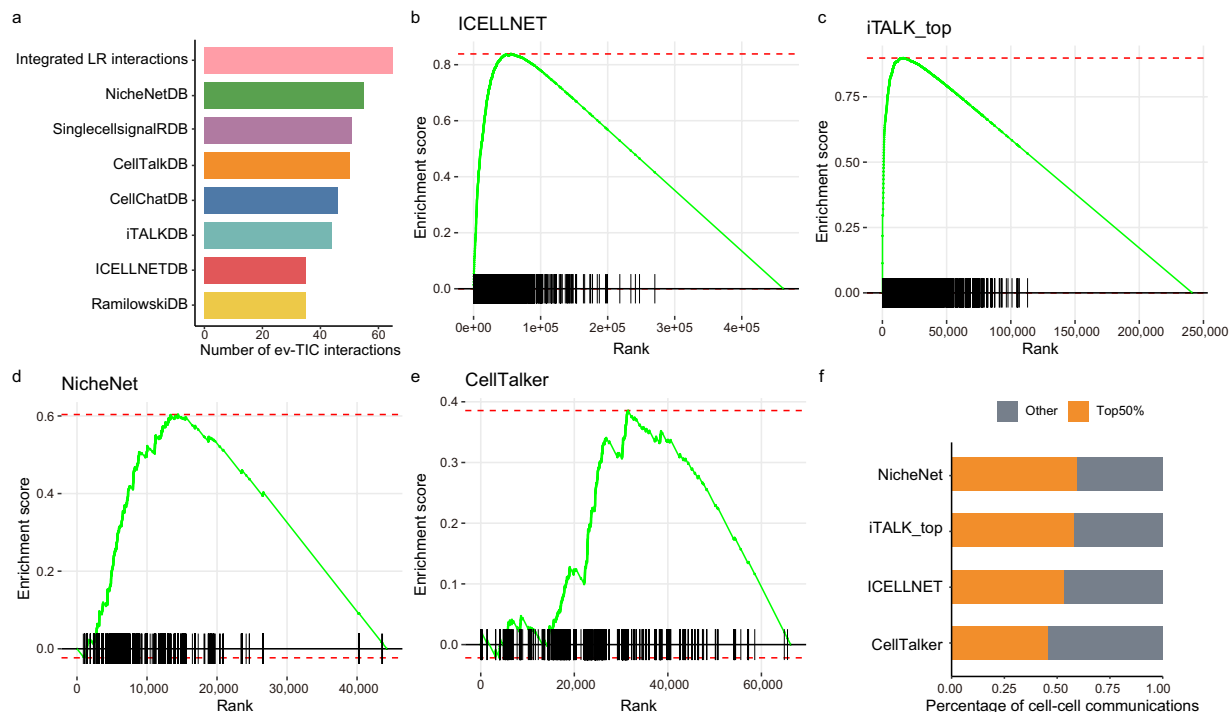
Fig. 4 The improvement of TICCom for prediction accuracy. (**a**) The number of experimentally-verified TIC (ev-TIC) communications in seven *human* resources and the integrated LR interactions. (**b-e**) The results of gene set enrichment analysis. To further illustrate the improvement of TICCom in prediction accuracy, we applied different algorithms, including ICELLNET, iTALK-top, NicheNet, and CellTalker, to infer cell-cell communication based on the integrated ligand-receptor interactions and iTALKDB LR interaction dataset in basal cell carcinoma (GSE123813[65]). The prediction result integrated from predicted cell-cell interactions inferred by these algorithms was used as the gene set. The prediction result of a single method was used as the ordered gene list, ranked by the communication score. (**f**) The percentage of top 50% of the cell-cell communication inferred from a single method occupied in the integrated result.

of experiments from Jan. 1993 to Jul. 2019. We proofread and validated the TIC communication by pulling low-throughput experiments from articles such as immunoprecipitation assays, qPCR, ELISPOT assays and Western blot assays. These interactions were divided into three categories: (1) 186 direct interactions, which meant these interactions directly occurred between tumor cells and immune cells; (2) 113 secretory interactions, which meant molecules derived from tumor cells or immune cells bound to corresponding receptors influencing TIC communication; and (3) 440 indirect interactions, which meant these interactions occurred within tumor cells or immune cells and were essential to tumor immunity (Fig. 2a). We carefully unified the names of cancer categories and immune cells based on the clinician's suggestions. Cancer categories are listed in Supplementary Table 2. These interactions involved 14 immune cells and 57 cancer subtypes of 23 cancer types (Fig. 2b,c, Supplementary Table 2). The number of TIC interactions varied across different tumors. This may be due to that the majority of experimentally-verified TIC interactions originate from these cancers, such as skin cancer, breast cancer, and lung cancer, in which tumor immunity has been a research hotspot, and cancer cell-immune cell crosstalk in other cancers may not have been explored in depth. In addition, the number of LR interactions varied across these seven *human* LR interaction datasets. Only 294 LR interactions were shared by seven *human* datasets (Fig. 2d). There were 1,157 common LR interactions in two *mouse* LR interaction datasets, accounting for 42% and 57% of the total, respectively (Fig. 2e). In order to obtain more comprehensive and precise resources on LR interactions, 14,190 *human* LR interactions and 3,650 *mouse* LR interactions were integrated from seven *human* datasets and two *mouse* datasets, respectively. We unified the gene symbols and Ensembl gene IDs of all the involved genes. The functions of these LR interactions were annotated by hand. In order to strengthen the credibility of LR interactions, these integrated interactions were grouped into the manually curated subclass or the predicted subclass based on their identification methods in previous studies. To guarantee the precision of interaction strength predicted by the Interaction Intensity module and TItalk provided by TICCom, we designed a p value that determined whether the real interaction strength was larger than the random one. The interaction strength was inferred based on bulk RNA-seq data from 25 cancer types (Fig. 3a,b, Supplementary Table 1). In addition, to ensure the accuracy of the predicted cell-cell crosstalk based on 32 scRNA-seq datasets from 13 cancer types (Fig. 3c, Table 1), we combined the results inferred from five algorithms in the Prediction module and used the integrated LR interactions as reference interactions. Before uploading these five datasets into TICCom, we re-checked the metadata of the MySQL database.

To improve the prediction accuracy of cell communication, two different aspects were considered by TICCom: the integration of ligand-receptor (LR) pair datasets and the integration of cell-cell interactions predicted based on different algorithms. We first assessed the contribution of integrating LR interaction datasets to prediction accuracy compared to a single LR resource. The result showed that the integrated data contained more experimentally-verified TIC interactions than single resources (Fig. 4a). On the other hand, we evaluated whether the integrated predicted cell-cell interactions tended to be optimized preferentially by each individual algorithm. We found that integrated predicted cell-cell interactions had higher communication scores in each algorithm by Gene Set Enrichment Analysis (GSEA) (Fig. 4b–e). Additionally, more than half of integrated predicted cell-cell communications accounted for the top 50% of results predicted by individual algorithms (Fig. 4f). The results indicated that integrating both LR interaction datasets and cell-cell interactions predicted by different algorithms could improve prediction accuracy at different levels.

Code availability

The R codes used to generate datasets in TICCom were shared on Github[36] (https://github.com/yunjinxie/TICCom-dataset) with the identifier (https://doi.org/10.5281/zenodo.8060109). All software tools used in this study are freely available.

References

1. Zhang, Y. & Zhang, Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cell Mol Immunol* **17**, 807–821, https://doi.org/10.1038/s41423-020-0488-6 (2020).
2. Zhang, Y. *et al.* Checkpoint therapeutic target database (CKTTD): the first comprehensive database for checkpoint targets and their modulators in cancer immunotherapy. *J Immunother Cancer* **8**, https://doi.org/10.1136/jitc-2020-001247 (2020).
3. Garner, H. & de Visser, K. E. Immune crosstalk in cancer progression and metastatic spread: a complex conversation. *Nat Rev Immunol* **20**, 483–497, https://doi.org/10.1038/s41577-019-0271-z (2020).
4. Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat Rev Genet* **17**, 441–458, https://doi.org/10.1038/nrg.2016.67 (2016).
5. Whiteside, T. L. The tumor microenvironment and its role in promoting tumor growth. *Oncogene* **27**, 5904–5912, https://doi.org/10.1038/onc.2008.271 (2008).
6. Andrews, N. *et al.* An unsupervised method for physical cell interaction profiling of complex tissues. *Nat Methods* **18**, 912–920, https://doi.org/10.1038/s41592-021-01196-2 (2021).
7. Cabello-Aguilar, S. *et al.* SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res* **48**, e55, https://doi.org/10.1093/nar/gkaa183 (2020).
8. Hara, T. *et al.* Interactions between cancer cells and immune cells drive transitions to mesenchymal-like states in glioblastoma. *Cancer Cell* **39**, 779–792 e711, https://doi.org/10.1016/j.ccell.2021.05.002 (2021).
9. Schurch, C., Riether, C., Matter, M. S., Tzankov, A. & Ochsenbein, A. F. CD27 signaling on chronic myelogenous leukemia stem cells activates Wnt target genes and promotes disease progression. *J Clin Invest* **122**, 624–638, https://doi.org/10.1172/JCI45977 (2012).
10. Dessarthe, B. *et al.* CRTAM receptor engagement by Necl-2 on tumor cells triggers cell death of activated Vgamma9Vdelta2 T cells. *J Immunol* **190**, 4868–4876, https://doi.org/10.4049/jimmunol.1202596 (2013).
11. Dong, H. *et al.* Tumor-associated B7-H1 promotes T-cell apoptosis: a potential mechanism of immune evasion. *Nat Med* **8**, 793–800, https://doi.org/10.1038/nm730 (2002).
12. Wischhusen, J. *et al.* Identification of CD70-mediated apoptosis of immune effector cells as a novel immune escape pathway of human glioblastoma. *Cancer Res* **62**, 2592–2599 (2002).
13. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* **15**, 1484–1506, https://doi.org/10.1038/s41596-020-0292-x (2020).
14. Wang, Y. *et al.* iTALK: an R Package to Characterize and Illustrate Intercellular Communication. Preprint at https://www.biorxiv.org/content/10.1101/507871v1 (2019).
15. Noel, F. *et al.* Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat Commun* **12**, 1089, https://doi.org/10.1038/s41467-021-21244-x (2021).
16. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12**, 1088, https://doi.org/10.1038/s41467-021-21246-9 (2021).
17. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* **17**, 159–162, https://doi.org/10.1038/s41592-019-0667-5 (2020).
18. Ramilowski, J. A. *et al.* A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun* **6**, 7866, https://doi.org/10.1038/ncomms8866 (2015).
19. Shao, X. *et al.* CellTalkDB: a manually curated database of ligand-receptor interactions in humans and mice. *Brief Bioinform* **22**, https://doi.org/10.1093/bib/bbaa269 (2021).
20. Baccin, C. *et al.* Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat Cell Biol* **22**, 38–48, https://doi.org/10.1038/s41556-019-0439-6 (2020).
21. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740, https://doi.org/10.1093/bioinformatics/btr260 (2011).
22. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat Biotechnol* **37**, 367–369, https://doi.org/10.1038/s41587-019-0055-9 (2019).
23. Papatheodorou, I. *et al.* Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res* **46**, D246–D251, https://doi.org/10.1093/nar/gkx1158 (2018).
24. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* **41**, D991–995, https://doi.org/10.1093/nar/gks1193 (2013).
25. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210, https://doi.org/10.1093/nar/30.1.207 (2002).
26. Sun, D. *et al.* TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res* **49**, D1420–D1430, https://doi.org/10.1093/nar/gkaa1020 (2021).
27. Hong, F. *et al.* Single-Cell Analysis of the Pan-Cancer Immune Microenvironment and scTIME Portal. *Cancer Immunol Res* **9**, 939–951, https://doi.org/10.1158/2326-6066.CIR-20-1026 (2021).

28. Zheng, L. L. *et al*. ColorCells: a database of expression, classification and functions of lncRNAs in single cells. *Brief Bioinform* **22**, https://doi.org/10.1093/bib/bbaa325 (2021).
29. Yuan, Z. *et al*. SODB facilitates comprehensive exploration of spatial omics data. *Nat Methods* **20**, 387–399, https://doi.org/10.1038/s41592-023-01773-7 (2023).
30. Cillo, A. R. *et al*. Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity* **52**, 183–199 e189, https://doi.org/10.1016/j.immuni.2019.11.014 (2020).
31. Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* **42**, D92–97, https://doi.org/10.1093/nar/gkt1248 (2014).
32. Han, H. *et al*. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* **46**, D380–D386, https://doi.org/10.1093/nar/gkx1013 (2018).
33. Bovolenta, L. A., Acencio, M. L. & Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* **13**, 405, https://doi.org/10.1186/1471-2164-13-405 (2012).
34. Vafaee, F. *et al*. ORTI: An Open-Access Repository of Transcriptional Interactions for Interrogating Mammalian Gene Expression Data. *PLoS One* **11**, e0164535, https://doi.org/10.1371/journal.pone.0164535 (2016).
35. Xie, Y. *et al*. A global database for modeling tumor-immune cell communication. *figshare* https://doi.org/10.6084/m9.figshare.22578031.v2 (2023).
36. Yunjinxie, yunjinxie/TICCom-dataset: The codes used to generate datasets in TICCom. *Zenodo*, https://doi.org/10.5281/zenodo.8060109 (2023).
37. Chung, W. *et al*. GEO. https://identifiers.org/geo/GSE145137 (2020).
38. Caron, M. *et al*. GEO. https://identifiers.org/geo/GSE132509 (2020).
39. Di Genua, C., Thongjuea, S. & Nerlov, C. GEO. https://identifiers.org/geo/GSE142213 (2020).
40. van Galen, P., Hovestadt, V. & Bernstein, B. GEO. https://identifiers.org/geo/GSE116256 (2019).
41. Wang, L., Dai, J., Han, R. & Jin, W. GEO. https://identifiers.org/geo/GSE143423 (2020).
42. Chung, W., Eum, H., Lee, H., Han, W. & Park, W. GEO. https://identifiers.org/geo/GSE75688 (2016).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP114962 (2018).
44. Tirosh, I., Hovestadt, V., Suva, M. & Filbin, M. GEO. https://identifiers.org/geo/GSE102130 (2018).
45. Yuan, J., Sims, P. & Levitin, H. GEO. https://identifiers.org/geo/GSE103224 (2018).
46. Diaz, A. *et al*. GEO. https://identifiers.org/geo/GSE138794 (2019).
47. Wang, R. *et al*. GEO. https://identifiers.org/geo/GSE139448 (2020).
48. Wang, L., Catalan, F., Babikir, H., Shamardani, K. & Diaz, A. GEO. https://identifiers.org/geo/GSE141982 (2020).
49. Tirosh, I., Venteicher, A., Suva, M. & Regev, A. GEO. https://identifiers.org/geo/GSE70630 (2016).
50. Darmanis, S. & Quake, S. GEO. https://identifiers.org/geo/GSE84465 (2017).
51. Tirosh, I., Venteicher, A., Suva, M. & Regev, A. GEO. https://identifiers.org/geo/GSE89567 (2017).
52. Tirosh, I., Puram, S. & Parikh, A. GEO. https://identifiers.org/geo/GSE103322 (2017).
53. Zhang, L. *et al*. GEO. https://identifiers.org/geo/GSE146771 (2020).
54. Wang, X. GEO. https://identifiers.org/geo/GSE125449 (2019).
55. Ahn, M. & Lee, H. GEO. https://identifiers.org/geo/GSE131907 (2020).
56. Lambrechts, D. *et al*. Phenotype molding of stromal cells in the lung tumor microenvironment. *ArrayExpress* https://identifiers.org/arrayexpress:E-MTAB-6149 (2018).
57. Zhang, W., Forbes, E., Jin, G. & Song, Q. GEO. https://identifiers.org/geo/GSE117570 (2019).
58. Zilionis, R. *et al*. GEO. https://identifiers.org/geo/GSE127465 (2019).
59. Rao, M. *et al*. GEO. https://identifiers.org/geo/GSE140312 (2020).
60. Shih, A. & Lee, A. GEO. https://identifiers.org/geo/GSE118828 (2018).
61. *Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA001160 (2019).
62. Moncada, R. GEO. https://identifiers.org/geo/GSE111672 (2020).
63. Jerby-Arnon, L. *et al*. GEO. https://identifiers.org/geo/GSE115978 (2018).
64. Tirosh, I. & Izar, B. GEO. https://identifiers.org/geo/GSE72056 (2016).
65. Yost, K. *et al*. GEO. https://identifiers.org/geo/GSE123813 (2019).
66. Paulson, K. & Chapuis, A. GEO. https://identifiers.org/geo/GSE117988 (2018).
67. Durante, M., Rodriguez, D. & Kurtenbach, S. GEO. https://identifiers.org/geo/GSE139829 (2020).
68. Li, S. & Zhang, P. GEO. https://identifiers.org/geo/GSE134520 (2019).

Acknowledgements

Author contributions

Yunjin Xie, Jingyi Shi, and Mengjia Xu performed data curation; Weiwei Zhou and Donghao Li constructed the database and designed the web interface; Yunjin Xie, Jingyi Shi, and Jianing Li performed data analysis; Zijing Lin unified the names of cancer categories; Tingting Shao and Jingyi Shi designed the algorithms; Yunjin Xie wrote the manuscript; Tingting Shao and Juan Xu provided funding for the project; Juan Xu, Tingting Shao and Shujun Cheng conceived of the project and provided critical advice on the study.

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/s41597-023-02342-5.

Correspondence and requests for materials should be addressed to S.C., T.S. or J.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.