Database, 2020, 1–6
doi: 10.1093/database/baaa020
Original article

Original article

ctcRbase: the gene expression database of circulating tumor cells and microemboli

Lei Zhao[1,†], Xiaohong Wu[2,†], Tong Li[3,†], Jian Luo[1,*] and Dong Dong[1,4,5,*]

[1]Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, No.500 Dongchuan Road, Shanghai, 200241 China, [2]Department of General Surgery, the Affiliated Yixing Hospital of Jiangsu University, No. 75 Zhenguan Road, Yixing, Jiangsu 214200, China, [3]Thyroid and breast surgery, the Fourth Hospital of Jinan City, No. 50 Shifan Road, Jinan, Shandong 250021, China, [4]Cancer Institute, Xuzhou Medical University, No. 84 West huaihai Road, Xuzhou, Jiangsu 221006, China and [5]Center of Clinical Oncology, Affiliated Hospital of Xuzhou Medical University, No.315 West huaihai Road, Xuzhou, Jiangsu 221006, China

*Corresponding author: Email: ddong@xzhmu.edu.cn
Correspondence may also be addressed to Jian Luo. Email: jluo@bio.ecnu.edu.cn
†These authors contributed equally to this work

Citation details: Zhao,L., Wu,X., Li,T. *et al.* ctcRbase: the gene expression database of circulating tumor cells and microemboli. *Database* (2020) Vol. 2020: article ID baaa020; doi:10.1093/database/baaa020

Received 21 November 2019; Revised 16 February 2020; Accepted 17 February 2020

Abstract

Circulating tumor cells/microemboli (CTCs/CTMs) are malignant cells that depart from cancerous lesions and shed into the bloodstream. Analysis of CTCs can allow the investigation of tumor cell biomarker expression from a non-invasive liquid biopsy. To date, high-throughput technologies have become a powerful tool to provide a genome-wide view of transcriptomic changes associated with CTCs/CTMs. These data provided us much information to understand the tumor heterogeneity, and the underlying molecular mechanism of tumor metastases. Unfortunately, these data have been deposited into various repositories, and a uniform resource for the cancer metastasis is still unavailable. To this end, we integrated previously published transcriptome datasets of CTCs/CTMs and constructed a web-accessible database. The first release of ctcRbase contains 526 CTCs/CTM samples across seven cancer types. The expression of 14 631 mRNAs and 3642 long non-coding RNAs of CTCs/CTMs were included. Experimental validations from the published literature are also included. Since CTCs/CTMs are considered to be precursors of metastases, ctcRbase also collected the expression data of primary tumors and metastases, which allows user to discover a unique 'circulating tumor cell gene signature' that is distinct from primary tumor and metastases. An easy-to-use database was constructed to query and browse CTCs/CTMs genes. ctcRbase can be freely accessible at http://www.origin-gene.cn/database/ctcRbase/.

Introduction

Circulating tumor cells (CTCs) are tumor cells that originate from either primary or metastatic tumors and travel through systemic circulation to distant organs, where they can initiate metastatic lesions (1–4). Circulating tumor microemboli (CTMs) are clusters of CTCs that can play an important role in metastatic cascade (5). Due to the intratumor heterogeneity, traditional surgical biopsies that are taken from one part of a tumor always miss information contained in other active regions. In contrast, CTCs/CTMs can consist of a mixture of cells shed from multiple active tumor regions, potentially providing a better representation of the invasive clones (1, 6). Because of the minimally invasive procedure of isolating CTCs/CTMs, it is a more practical approach for repeatedly monitoring disease progression (7). Analysis of CTCs allows investigation of cancer cell biomarker expression from a non-invasive liquid biopsy, and their analysis has emerged as one of the hottest fields in cancer research (8–11). CTCs have exhibited potential for tumor diagnosis, treatment and monitoring.

Over the past years, great efforts have been made to detect and separate CTCs. The classical method of positive selection is utilizing epithelial cell adhesion molecule (EpCAM), which is consistently expressed by epithelial-derived tumor cells and the absence from the normal leukocytes (12). The microfluidic CTC isolation technologies can effectively deplete leukocytes without manipulating CTCs (13). It preserves cell viability and ensures the high-quality RNA content to the greatest extent. High-throughput technology has become a powerful tool to provide a genome-wide view of transcriptomic changes associated with CTCs/CTMs (14–18). These data provided us much information to understand the tumor heterogeneity, and the underlying mechanism of tumor metastases at the single-cell perspective. Single-cell sequencing provides a new method to identify etiology at the whole-genome wide level (19, 20). After CTC isolation, single-cell sequencing can be applied to identify genomic and transcriptomic characteristics of CTCs.

Unfortunately, these data have been deposited in various repositories, and a uniform resource for the cancer metastasis is still unavailable. To this end, we integrated previously published transcriptome datasets of CTCs/CTMs and constructed a web-accessible database. An easy-to-use interface was constructed to query and browse CTC/CTM genes. Current version of ctcRbase is not only a comprehensive, update-to-date resource of CTC/CTM gene expression data sets, but also can compare CTC/CTM expression to tumor/WBC expression for the seven cancers included. In addition, all expression data of CTCs/CTM contained in ctcRbase can be downloaded and reprocessed by users, which enhances the utility of ctcRbase. This knowledge would be helpful for researchers to better understand the molecular mechanisms underlying tumor metastasis, relapse and chemoresistance and might eventually aid in the development of new targeted cancer therapies.

Database content and usage

Data summary

We curated RNA-seq raw data of CTCs/CTMs from the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) or Short Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra/) using the keywords ‘CTC’ or ‘CTCs’ or ‘Circulating Tumor Cells’ or ‘CTM’ or ‘Circulating Tumor Microemboli’ or ‘CTC cluster’ and ‘RNA-seq’. A total of 526 samples from 17 datasets were downloaded (Table 1) across seven cancer types, including breast cancer (BRCA), colorectal cancer (COAD), prostate cancer (PRAD), non-small cell lung cancer (LUSC), pancreatic cancer (PAAD), melanoma (SKCM) and liver cancer (LIHC) (Figure 1A).

Here, ctcRbase also provides the expression profiles of primary tumors and metastasis sites. The expression data (FPKM values) of primary tumors were download from The Cancer Genome Atlas (TCGA). We manually curated the TCGA clinical data and removed all the metastasis samples. Tumor metastasis expression data were collected from the Human Cancer Metastasis Database (HCMDB) (21). For tumor metastasis, a total of 556 samples from 18 metastasis sites were downloaded.

Analysis of RNA-seq data

The sequencing raw data were first trimmed by removing adapters using TrimGalore. Then all the CTC/CTM data were processed through a consistent pipeline (Figure 1B). Sequencing reads were mapped to the human reference genome (GRCh38) by bowtie (22), and then RSEM software (23) was used to calculate the read counts and FPKM (Fragments Per Kilobase Per Million). In total, 14 631 mRNAs and 3642 lncRNAs were collected in ctcRbase. The functions of some genes in CTCs/CTMs have been studied in previous works. To better annotate those genes expressed in CTCs/CTMs, we searched Pubmed database using the keywords of ‘CTC’ or ‘CTCs’ or ‘Circulating Tumor Cells’ or ‘CTM’ or ‘Circulating Tumor Microemboli’ or ‘CTC cluster’. After all the papers were downloaded, we filtered those papers which have no detailed functions of specific genes. Then, we screened out those papers and extracted the descriptions of conclusions about the genes expressed in CTCs/CTMs. Finally, we

Table 1. Data statistics

Cancer type	CTC sample numbers	Metastasis sample numbers	CTCs/CTM isolation method
Breast cancer	339	101	CTC-iChip, micromanipulation & immunofluorescence, microfluidics & immunofluorescence, immunomagnetic & FACS, microfluidics & immunofluorescence
Colorectal cancer	18	50	Microfiltration & immunofluorescence
Melanoma	6	236	Immunomagnetic purification
Non-small cell lung cancer	10	5	CTC-iChip
Pancreatic cancer	19	59	CTC-iChip
Prostate cancer	89	85	CTC-iChip
Liver cancer	45	20	CTC-iChip & immunofluorescence

The sample numbers of CTCs/CTM from GEO/SRA and metastasis from HCMDB were provided. Furthermore, the isolation methods of CTCs/CTM in different cancer types were provided

obtained the conclusions including gene expression levels, biological implications and DNA mutation information of genes in 320 papers.

Competing endogenous RNA network prediction

lncRNAs play important roles in tumorigenesis and metastasis (24, 25). Recent studies have reported that diverse RNAs could communicate with each other and have reciprocal regulation by acting as competing endogenous RNAs. To understand this complex RNA crosstalk in CTCs/CTMs, competing endogenous RNA networks were predicted. The mRNA-lncRNA *cis*-regulatory relationships were defined as pairs consisting of genes located within a genomic window of 100 kb. miRNA targets were predicted by RNAhybrid (26), miRanda (27) and PITA software (28). Those genes that were identified by at least two software were regarded as miRNA targets. Cytoscape software was used to visualize the network.

Database query and search platform

A user-friendly web interface was built to present the ctcRbase. Several ways were provided to allow database query. First, a search engine was developed in ctcRbase using gene names (Gene Symbol, EnsembleID or EntrezID) from the ‘Search’ page. Users can input their interested genes in the textbox (mRNA and lncRNA), and all the items that contain the query genes in the database can be derived. The result page of search function lists the database ID, gene symbol, category, cancer type and CTC/CTM isolation method. Additionally, users can select mRNA or lncRNA category from the ‘Browse’ page (Figure 1C).

Database implementation

A user-friendly web interface was developed to present ctcRbase. All data was deposited into MySQL database. The web interface for searching and browsing was implemented by JavaScript and PHP.

The result page for gene search

The result page for single gene search has five major sections: (i) general information; (ii) gene information; (iii) literature description; (iv) gene expressions; and (v) competing endogenous RNA network (Figure 1C). General information provides the Database ID, cancer type, dataset and CTC/CTM isolation methods. Gene information shows the gene annotation information, including gene symbol, full name, category, synonyms, location and gene summary. In the literature description section, ctcRbase provides the literature descriptions curated from Pubmed about the functional implications of specific gene in CTCs/CTMs. In the gene expression section, ctcRbase shows the expression levels of primary cancer, CTCs/CTM and whole blood cells (WBC). Moreover, ctcRbase allows users to compare the gene expression among primary tumors, CTCs/CTMs and metastasis. The last section provides the competing endogenous RNA network. Users can online visualize the regulatory network of miRNA–lncRNA–mRNA.

Perspectives and concluding remarks

To date, our attention is turning to the non-invasive liquid biopsies, which enables analysis of CTCs/CTMs in bodily fluids. CTCs/CTMs have been proved to be a promising approach in cancer diagnostics, with the applications ranging from tumor early detection to treatment selection (1). Comparing to other tumor components, such as circulating tumor DNA (ctDNA), CTCs/CTMs provide a source of intact genomic and transcriptomic information for lineage-based analysis. Recently, single- or bulk-cell sequencing data of CTCs/CTMs have been available, which provides us great opportunities to detect the transcriptomic landscape

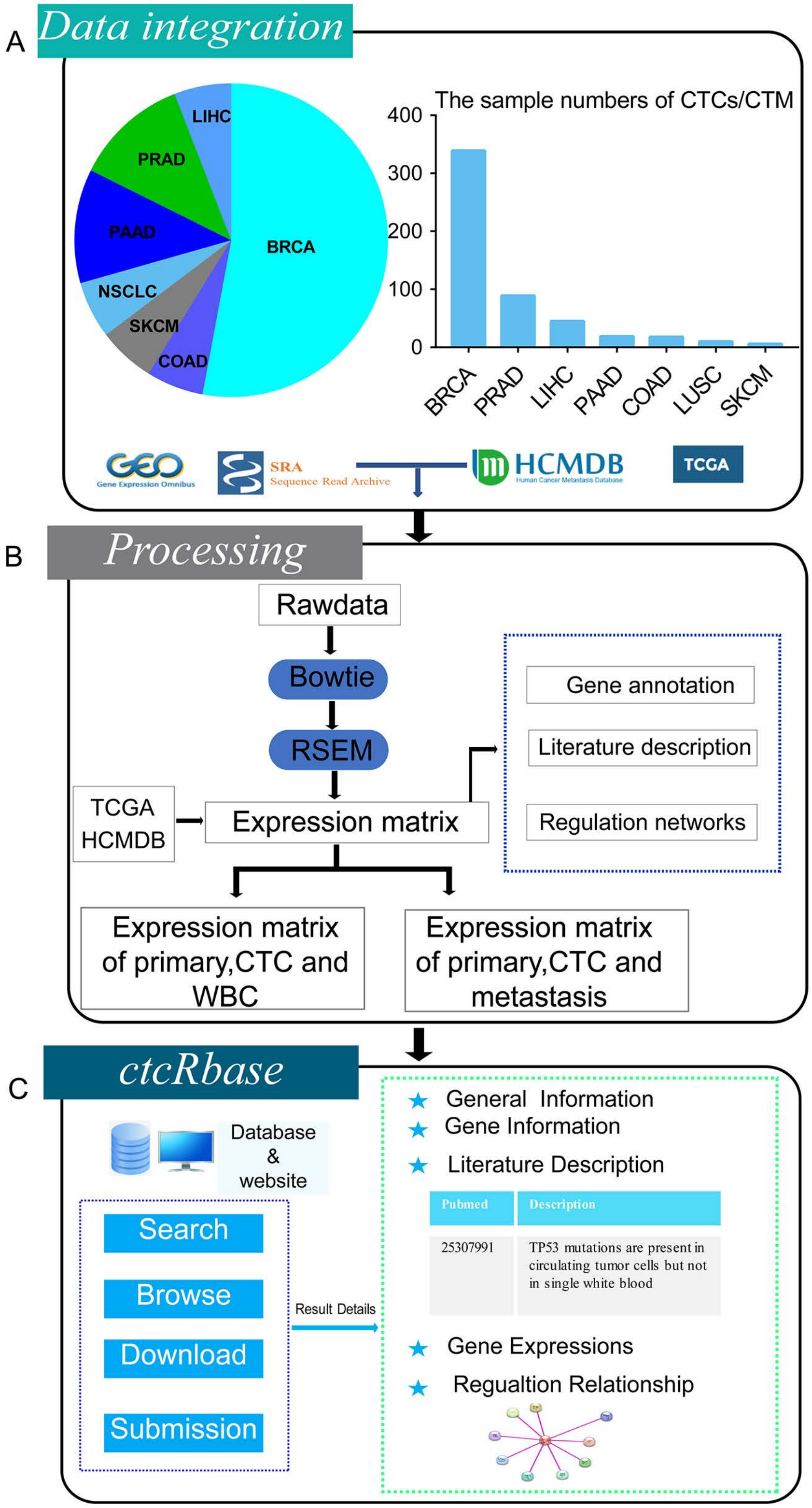

Figure 1. Overall design of ctcRbase. (**A**) Data distribution of CTCs/CTMs in ctcRbase. ctcRbase contains CTC/CTM data across seven cancer types, including breast cancer (BRCA), colorectal cancer (COAD), melanoma (SKCM), non-small cell lung cancer (LUSC), pancreatic cancer (PANC), prostate cancer (PRAD) and liver cancer (LIHC). The data from TCGA and HCMDB was used to compare the expression of WBC, primary tumor, CTCs/CTM and metastasis sites. (**B**) Data processing of ctcRbase. Bowtie and RSEM was used to get expression matrix. (**C**) The main functions of ctcRbase include 'Search', 'Browse', 'Download' and 'Submission'. Result page contains five parts: (i) general information; (ii) gene information; (iii) literature description; (iv) gene expressions; and (v) competing endogenous RNA network.

of CTCs/CTMs and can provide RNA-based signatures enabling high specificity detection of CTCs/CTMs (29, 30). Furthermore, CTCs has been regarded as the seeds for the subsequent metastases in distant organs (31). These data provide us a comprehensive understanding of the metastatic cascade.

Although many single-cell RNA sequencing data have been published, it is hard for researchers to use these data and there has no accessible database. ctcRbase is the first database for searching and visualizing the expression pattern of CTCs/CTMs and their primary tumor, WBC and metastasis sites. It provides a better way for researchers to understand the gene expression pattern in tumor metastasis. The data can reveal the gene regulation process in metastasis and can facilitate our understanding of the metastasis hallmark. ctcRbase will be continuously updated and provide more information including: (i) more upcoming CTCs/CTM transcriptome data; (ii) DNA methylation data of CTCs/CTM and their paired primary tumor; and (iii) more comprehensive CTCs/CTM genetic and genomic data, including copy number variation data and mutation variation data. We expected that ctcRbase can contribute to researchers' understanding about tumor metastasis mechanism and even facilitate to find the therapeutic method of tumor in the future.

Authors' contributions

L.Z. and D.D. performed the most analyses and drafted the paper. L.Z., X.W., T.L. and J.L. set up the database and participated in analyses. L.Z. and D.D. participated in the data curation. All authors read and approved the final manuscript.

Funding

Jiangsu Provincial Key Medical Discipline; The Project of Invigorating Health Care through Science, Technology and Education (No. ZDXKA2016014).

Conflict of interest. None declared.

References

1. Alix-Panabieres,C. and Pantel,K. (2013) Circulating tumor cells: liquid biopsy of cancer. *Clin Chem*, **59**, 110–118.
2. Cristofanilli,M., Budd,G.T., Ellis,M.J. *et al.* (2004) Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N Engl J Med*, **351**, 781–791.
3. Aguirre-Ghiso,J.A., Bragado,P. and Sosa,M.S. (2013) Metastasis awakening: targeting dormant cancer. *Nat Med*, **19**, 276–277.
4. Aceto,N., Bardia,A., Miyamoto,D.T. *et al.* (2014) Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell*, **158**, 1110–1122.
5. Hou,J.M., Krebs,M.G., Lancashire,L. *et al.* (2012) Clinical significance and molecular characteristics of circulating tumor cells and circulating tumor microemboli in patients with small-cell lung cancer. *J Clin Oncol*, **30**, 525–532.
6. Rossi,G., Mu,Z., Rademaker,A.W. *et al.* (2018) Cell-free DNA and circulating tumor cells: comprehensive liquid biopsy analysis in advanced breast cancer. *Clin Cancer Res*, **24**, 560–568.
7. Yu,M., Stott,S., Toner,M. *et al.* (2011) Circulating tumor cells: approaches to isolation and characterization. *J Cell Biol*, **192**, 373–382.
8. Li,J., Han,X., Yu,X. *et al.* (2018) Clinical applications of liquid biopsy as prognostic and predictive biomarkers in hepatocellular carcinoma: circulating tumor cells and circulating tumor DNA. *J Exp Clin Cancer Res*, **37**, 213.
9. Ye,Q., Ling,S., Zheng,S. and Xu,X. (2019) Liquid biopsy in hepatocellular carcinoma: circulating tumor cells and circulating tumor DNA. *Mol Cancer*, **18**, 114.
10. Zainfeld,D. and Goldkorn,A. (2018) Liquid biopsy in prostate cancer: circulating tumor cells and beyond. *Cancer Treat Res*, **175**, 87–104.
11. Zhang,X., Ju,S., Wang,X. and Cong,H. (2019) Advances in liquid biopsy using circulating tumor cells and circulating cell-free tumor DNA for detection and monitoring of breast cancer. *Clin Exp Med*, **19**, 271–279.
12. Allard,W.J., Matera,J., Miller,M.C. *et al.* (2004) Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clin Cancer Res*, **10**, 6897–6904.
13. Ozkumur,E., Shah,A.M., Ciciliano,J.C. *et al.* (2013) Inertial focusing for tumor antigen-dependent and -independent sorting of rare circulating tumor cells. *Sci Transl Med*, **5**, 179ra47.
14. Wang,Y., Wu,N., Liu,J. *et al.* (2015) FusionCancer: a database of cancer fusion genes derived from RNA-seq data. *Diagn Pathol*, **10**, 131.
15. Miyamoto,D.T., Zheng,Y., Wittner,B.S. *et al.* (2015) RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science*, **349**, 1351–1356.
16. Bhan,I., Mosesso,K., Goyal,L. *et al.* (2018) Detection and analysis of circulating epithelial cells in liquid biopsies from patients with liver disease. *Gastroenterology*, **155**, 2016–2018 e11.
17. Boral,D., Vishnoi,M., Liu,H.N. *et al.* (2017) Molecular characterization of breast cancer CTCs associated with brain metastasis. *Nat Commun*, **8**, 196.
18. Yu,M., Ting,D.T., Stott,S.L. *et al.* (2012) RNA sequencing of pancreatic circulating tumour cells implicates WNT signalling in metastasis. *Nature*, **487**, 510–513.
19. Ting,D.T., Wittner,B.S., Ligorio,M. *et al.* (2014) Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep*, **8**, 1905–1918.
20. Lapin,M., Tjensvoll,K., Oltedal,S. *et al.* (2017) Single-cell mRNA profiling reveals transcriptional heterogeneity among pancreatic circulating tumour cells. *BMC Cancer*, **17**, 390.

21. Zheng,G., Ma,Y., Zou,Y. *et al.* (2018) HCMDB: the human cancer metastasis database. *Nucleic Acids Res*, **46**, D950–D955.
22. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.
23. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
24. Carninci,P., Kasukawa,T., Katayama,S. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
25. Prensner,J.R. and Chinnaiyan,A.M. (2011) The emergence of lncRNAs in cancer biology. *Cancer Discov*, **1**, 391–407.
26. Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res*, **34**, W451–W454.
27. John,B., Enright,A.J., Aravin,A. *et al.* (2004) Human MicroRNA targets. *PLoS Biol*, **2**, e363.
28. Kertesz,M., Iovino,N., Unnerstall,U. *et al.* (2007) The role of site accessibility in microRNA target recognition. *Nat Genet*, **39**, 1278–1284.
29. Miyamoto,D.T., Lee,R.J., Kalinich,M. *et al.* (2018) An RNA-based digital circulating tumor cell signature is predictive of drug response and early dissemination in prostate cancer. *Cancer Discov*, **8**, 288–303.
30. Kalinich,M., Bhan,I., Kwan,T.T. *et al.* (2017) An RNA-based signature enables high specificity detection of circulating tumor cells in hepatocellular carcinoma. *Proc Natl Acad Sci U S A*, **114**, 1123–1128.
31. Ramaswamy,S., Ross,K.N., Lander,E.S. and Golub,T.R. (2003) A molecular signature of metastasis in primary solid tumors. *Nat Genet*, **33**, 49–54.