Bioinformatics, 2021, 1–3
doi: 10.1093/bioinformatics/btab405
Applications Note

Database and ontologies

Tfcancer: a manually curated database of transcription factors associated with human cancers

Qingqing Huang[†], Zhengtang Tan[†], Yanjing Li, Wenzhu Wang, Mei Lang, Changying Li
AQ3 AQ2 **and Zhiyun Guo ***

AQ4 School of Life Sciences and Engineering, Southwest Jiaotong University, Chengdu 610031, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor:: Christina Kendziorski
Received on January 21, 2021; revised on April 19, 2021; editorial decision on April 26, 2021; accepted on May 25, 2021

Abstract

Summary: Transcription factors (TFs) are critical regulation elements and its dysregulation can lead to a variety of cancers. However, currently, there are no such online resources for large-scale collection, storage and analysis of TF-cancer associations in those cancers. To fill this gap, we present a database called TFcancer (http://lcbb.swjtu.edu.cn/tfcancer/), which contains 3136 experimentally supported associations between 364 TFs and 33 TCGA cancers by manually curating more than 1800 literature. TFcancer mainly concentrates on four aspects: TF expression, molecular alteration, regulatory relationships between TFs and target genes, and biological processes and signaling pathways of TFs in cancers. TFcancer not only provides a user-friendly interface for browsing and searching but also allows flexible data downloading and user data submitting. It is believed that TFcancer is a helpful and valuable resource for researchers who seek to understand the functions and molecular mechanisms of TFs involved in human cancers.

Availability and implementation: The TFcancer are freely available at http://lcbb.swjtu.edu.cn/tfcancer/.

Contact: zhiyunguo@swjtu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Nowadays, thousands of experiments have been performed to clarify the functions of transcription factors (TFs) in tumorigenesis and cancer treatment, especially after the completion of the TCGA project. However, relevant information on TFs is fragmented and hides behind almost countless literatures, making it hard for researchers to systematically obtain and analyse the relationships between TFs and kinds of cancer types.

To date, several TF databases have been developed to explore this information. For example, TF-target gene regulation databases like TRRUST (Han *et al.*, 2018), ChIPSummitDB (Czipa *et al.*, 2020), hTFtarget (Zhang *et al.*, 2020), SEGreg (Tang *et al.*, 2019) and FFLtool (Xie *et al.*, 2020) and databases containing annotation and prediction of TFs such as REGULATOR (Wang and Nishida, 2015) and AnimalTFDB (Hu *et al.*, 2019). Nevertheless, almost all of them focus on the identification of TF–target interactions or TFs, but the relationship between TFs and cancers is rarely involved. Furthermore, most of these databases are mainly obtained through analysis of DNA sequence or high-throughput data, which can cause a problem as it may produce a lot of false-positive results (Garcia-Alonso *et al.*, 2019). To our knowledge, currently, there are no such online resources for large-scale collection of TF-cancer information based on manually curated literature.

Here, we present TFcancer (http://lcbb.swjtu.edu.cn/tfcancer/), a comprehensive database which manually curates thousands of literature involved in the relationships between TFs and 33 cancers types of TCGA. TFcancer mainly focuses on the influence of TFs on four aspects of cancer: differential expression, molecular alteration of TFs, the regulatory relationship between TFs and genes and the biological processes/signaling pathways affected by TFs in cancers. In this way, we obtained 3136 experimentally supported contexts of 364 TFs and 33 cancers by manually curating 1843 experimental literature published on PubMed. It is believed that TFcancer will facilitate studies to establish relationships between biological functions of TFs and cancers.

2 Materials and methods

2.1 Collection of literatures

To collect comprehensive information about TF functions in 33 cancers, we searched 'title/abstract' field of PubMed using the combination of the phrase 'TF' and 33 cancer names. It was noticed that

Table 1. Rules for scoring literatures

Item	TF name	TF	Cancer name	Others	Minus word	Sum score for each item
TI	18	14	12	8	–8	Sti
AB	8	6	4	2	–2	Sab
(for every sentence)[a]	Sab_tfs		Sab_rest			
MH/RN/OT	4	2	1	1	–2	Smh

[a]For each sentence in AB, the total score of TF related feature words was calculated as Sab_tfs and the total score of rest feature words was calculated as Sab_rest. To make sure the functions described in the sentence were regarded with TFs, we use Sab = Sab_tfs * Sab_rest when Sab_tfs > 0 and Sab_rest > 0. Otherwise we use Sab = Sab_rest.+ Sab_tfs.

there were varieties of aliases for cancers in thousands of literatures. To solve this issue, we obtained all aliases of those cancers through searching the MalaCards database which offered comprehensive annotation information about human diseases. For example, the final search keywords for UVM was '(TF*[Title/Abstract]) AND (Choroidal Melanoma*[Title/Abstract] OR Uveal Melanoma*[Title/Abstract] OR Malignant Melanoma* of Choroid[Title/Abstract] OR Malignant Melanoma* of Iris[Title/Abstract] OR Melanoma* of Uvea[Title/Abstract] OR Iris Melanoma*[Title/Abstract])'. The use of '*' ensured that both plural and singular forms of words were included.

2.2 Filter irrelevant literatures by scoring the retrieved literatures

Since there was a large quantity of literature involving 33 cancers in TF, and most of them were irrelevant, we designed a scoring system to filter out the unconcerned ones based on the occurrences of 'feature words'. 'Feature words' were those representative phrases or words which occurred most frequently in the four research fields mentioned above including TF gene names, corresponding cancer names, expression and biological terms of cancer, etc. (Table 1), and negative score words that were obviously not related to the four types of research (referred as 'minus words' in Table 1).

The following scoring rules were used to rank the relevance level of literature about TF functions and cancers. Firstly, the position of a word in literature reflected its relevance to the topic. PubMed provides MEDLINE format download which is a tagged field format displaying all fields of the MEDLINE record containing not only abstract (AB) and title (TI) but also other terms (OT) and MeSH Terms (MH). Specifically, feature words in TI were endowed with the highest score, the ones in AB with the second and the ones in OT and MH with the lowest (Table 1). Secondly, the frequency of feature word in literature also reflected their relevance. The more a feature word appeared in the article, the more likely this article belonged to the four research fields that we focused on. Therefore, the total score of a particular feature word in one field was obtained by multiplying the feature word score by the number of occurrences, and 25% was set as the literature filtering ratio to screen relevant literatures (Supplementary Fig. S1).

3 Results

3.1 Summaries of the TF-cancer associations

Finally, we screened 1843 articles and manually reviewed them and extracted a total of 3136 experimentally validated associations between 33 cancer types and 364 TFs. The associations mainly covered four aspects: (i) differential expression of TFs in cancers and normal tissues (e.g. high expression, low expression, dysregulation, etc.); (ii) molecular alterations such as SNP, rearrangement, CNV, promoter methylation, amplification, fusion genes and translocation of TFs; (iii) transcriptional regulatory interactions of TFs and other genes (e.g. TF target genes, TF targeted by genes, promoter binding, feedback/feed-forward loop, etc.) and (iv) cancer hallmarks, cancer-associated biological processes and pathways which were affected by TFs (e.g. proliferation, apoptosis, metastasis, invasion, endocrine resistance, etc.).

3.2 General web interface and database

TFcancer offers a user-friendly web interface to help users search, browse, download and submit associations between TFs and 33 cancers (Supplementary Fig. 2a). The search dropdown menu is applied with two modes: general search and advanced search. General search allows users to input different TF gene symbols/Ensembl ID and choose cancer types (Supplementary Fig. 2b). Advanced search is equipped with more options including cancer names, TF gene symbols, TF characteristics (high expression, amplification, translocation, etc.), regulation modes of TFs and target genes (positive, negative, feedback loop, etc.), interaction between TFs and genes and biological processes and pathways of TFs in cancers (Supplementary Fig. 2c).

After users submit their search requests, the preliminary result page will be shown. For each entry, there are eight columns containing cancer abbreviations, TF names, TF characteristics, genes that interact with TFs, regulation modes of the interactions between TFs and genes, cancer processes and pathways affected by TFs, PMID of literature and web link of 'Detail' page (Supplementary Fig. 2d). It is also convenient for users to search the results by directly inputting keywords at the top left corner of the page (Supplementary Fig. 2e). On the 'Detail' page of corresponding entry, the title of the literature and the original text of the association entry are accessible to users (Supplementary Fig. 1f). Besides, they can also browse TFcancer by clicking cancer type abbreviations on the 'Browse' page (Supplementary Fig. 2g). 'Download' page allows users to download TFcancer information of different cancer types (Supplementary Fig. 2h). Meanwhile, TFcancer welcomes users to submit novel experimentally supported TF-cancer associations on the 'Submit' page (Supplementary Fig. 2i).

3.3 Future development

The multiple species literature will be gathered in the next years. As more and more documents related to human TFs are published, we will continuously collect the latest data to maintain our database up-to-date.

Acknowledgements

The authors thank Fei Tang for supporting the early development of the TFcancer platform. They also thank Yin Zhang for testing the website.

Funding

This work was supported by the Basic Research Cultivation Support Program of Fundamental Research Funds for the Central Universities (2682021ZTPY016) and National Science and Technology Major Project of Infectious Diseases (2018ZX10101-003-001-008). AQ5

Conflict of Interest: none declared. AQ6 AQ7

References AQ8

Czipa,E. *et al.* (2020) ChIPSummitDB: a ChIP-seq-based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them. *Database (Oxford)*, **2020**, baz141.

Garcia-Alonso,L. *et al.* (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.*, **29**, 1363–1375.

Han,H. *et al.* (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.*, **46**, D380–D386.

Hu,H. *et al.* (2019) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.*, **47**, D33–D38.

Tang,Q. *et al.* (2019) SEGreg: a database for human specifically expressed genes and their regulations in cancer and normal tissue. *Brief. Bioinform.*, **20**, 1322–1328.

Wang,K. and Nishida,H. (2015) REGULATOR: a database of metazoan transcription factors and maternal factors for developmental studies. *BMC Bioinformatics*, **16**, 114.

Xie,G.Y. *et al.* (2020) FFLtool: a web server for transcription factor and miRNA feed forward loop analysis in human. *Bioinformatics*, **36**, 2605–2607.

Zhang,Q. *et al.* (2020) hTFtarget: a comprehensive database for regulations of human transcription factors and their targets. *Genom. Proteom. Bioinform.*, **18**, 120–128.

Author Query Form

Journal:	*Bioinformatics*
Article Doi:	10.1093/bioinformatics/btab405
Article Title:	Tfcancer: a manually curated database of transcription factors associated with human cancers
First Author:	Qingqing Huang
Corr. Author:	Zhiyun Guo

AUTHOR QUERIES – TO BE ANSWERED BY THE CORRESPONDING AUTHOR

These proofs are for checking purposes only. They are not in final publication format. Please do not distribute them in print or online. Do not publish this article, or any excerpts from it, anywhere else until the final version has been published with OUP. For further information, see https://academic.oup.com/journals/pages/authors

Figure resolution may be reduced in PDF proofs and in the online PDF, to manage the size of the file. Full-resolution figures will be used for print publication.

Select each question and describe any changes we should make on the proof. Changes against journal style will not be made and proofs will not be sent back for further editing.

AQ1: We will be using the running head "Tfcancer" during the print stage if necessary. It would be helpful if you could confirm the running head.

AQ2: Please check all author names and affiliations. Please check that author surnames have been identified by a pink background in the PDF version, and by green text in the html proofing tool version (if applicable). This is to ensure that forenames and surnames have been correctly tagged for online indexing.

AQ3: If your manuscript has figures or text from other sources, please ensure you have permission from the copyright holder. For any questions about permissions contact jnls.author.support@oup.com.

AQ4: Please provide department name and postal code for affiliation.

AQ5: Please check that funding is recorded in a separate funding section if applicable. Use the full official names of any funding bodies, and include any grant numbers.

AQ6: You may need to include a “conflict of interest” section. This would cover any situations that might raise any questions of bias in your work and in your article’s conclusions, implications, or opinions. Please see https://academic.oup.com/journals/pages/authors/authors_faqs/conflicts_of_interest.

AQ7: Please note that there will be a charge of £100/US$190 per page for extra printed pages above 8 pages for a Review, 7 pages for an Original Article, 4 pages for a Discovery Note and 2 pages for an Applications Note. If your article exceeds these lengths, please confirm that you accept the charge

AQ8: Journal policy requires authors to provide a data availability statement in their manuscript. Please confirm that this statement is included in your manuscript and that any required links or identifiers for your data are present in the manuscript as described or provide edits with the required information.