ONCOIMMUNOLOGY
2021, VOL. 10, NO. 1, e1944553 (12 pages)
https://doi.org/10.1080/2162402X.2021.1944553

Taylor & Francis
Taylor & Francis Group

ORIGINAL RESEARCH

OPEN ACCESS Check for updates

CanImmunother: a manually curated database for identification of cancer immunotherapies associating with biomarkers, targets, and clinical effects

Wenliang Zhang [a,b,c*,#], Binghui Zeng[d*], Huancai Lin[d], Wen Guan[c,e], Jing Mo[c], Song Wu[f], Yanjie Wei[b,g,h], Qianshen Zhang[a], Dongsheng Yu [d#], Weizhong Li [f,i,j#], and Godfrey Chi-Fung Chan [a,k#]

[a]Department of Pediatrics, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China; [b]Chinese Academy of Sciences, Shenzhen Institute of Advanced Technology, Shenzhen, Guangdong, China; [c]Department of Bioinformatics, Outstanding Biotechnology Co., Ltd.-Shenzhen, Shenzhen, China; [d]Guangdong Provincial Key Laboratory of Stomatology, Guanghua School of Stomatology, Hospital of Stomatology, Sun Yat-sen University, Guangzhou, China; [e]Guangdong Key Laboratory of Animal Conservation and Resource Utilization, Institute of Zoology, Guangdong Academy of Sciences, Guangzhou, China; [f]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China; [g]Center for High Performance Computing, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China; [h]CAS Key Laboratory of Health Informatics, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China; [i]Center for Precision Medicine, Sun Yat-sen University, Guangzhou, China; [j]Key Laboratory of Tropical Disease Control of Ministry of Education, Sun Yat-sen University, Guangzhou, China; [k]Department of Pediatrics and Adolescent Medicine, Faculty of Medicine, The University of Hong Kong, Hong Kong

ABSTRACT

As immunotherapy is evolving into an essential armamentarium against cancers, numerous translational studies associated with relevant biomarkers, targets, and clinical effects have been reported in recent years. However, a large amount of associated experimental data remains unexplored due to the difficulty in accessibility and utilization. Here, we established a comprehensive high-quality database for cancer immunotherapy called CanImmunother (http://www.biomedical-web.com/cancerit/) through manual curation on 4515 publications. CanImmunother contains 3267 experimentally validated associations between 218 cancer sub-types across 34 body parts and 484 immunotherapies with 642 biomarkers, 108 targets, and 121 control therapies. Each association was manually curated by professional curators, incorporated with valuable annotation and cross references, and assigned with an association score for prioritization. To help clinicians and researchers in identifying and discovering better cancer immunotherapy and their respective biomarkers and targets, CanImmunother offers user-friendly web applications including search, browse, excel table, association prioritization, and network visualization. CanImmunother presents a landscape of experimental cancer immunotherapy association data, serving as a useful resource to improve our insight and to facilitate further discovery of advanced immunotherapy options for cancer patients.

ARTICLE HISTORY
Received 25 December 2020
Revised 12 June 2021
Accepted 15 June 2021

KEYWORDS
Cancer immunotherapy; immune checkpoint; database; biomarker; tumor vaccine

Introduction

Cancer immunotherapy has emerged and rapidly developed over the past few decades[1]. Its basic principle is to promote and facilitate the immune system in targeting cancer cells. The technologies include single or bispecific monoclonal antibodies, immune checkpoint inhibitors, various forms of cellular therapies, immunomodulatory cytokines and chemokines, tumor vaccines, and so on [2]. One of the successful strategies is immune checkpoint inhibitor, which targets at the cytotoxic T lymphocyte-associated protein 4 (CTLA-4), the programmed cell death protein 1 (PD-1), or their ligands (such as PD-L1).[2,3] Currently, various forms of cancer immunotherapy have already been proven by clinical trials to be effective, and they become an essential part of contemporary cancer therapy [1]. Although applications of cancer immunotherapy cover a broad range of human cancers, their usefulness is restricted by whether particular cancer types have tumor-specific antigens or co-inhibitory molecules.[1,4] In addition, different forms of cancer immunotherapy have their own unique toxicity profiles, depending on their mechanism of action, which are distinct from the usual therapy related to toxicity encountered in chemotherapy.[1,5] However, useful association data about cancer immunotherapy and their predictive biomarkers, therapy combination, clinical efficacy, and adverse effects, may still be unexplored within the large amount of categorical literature, which are difficult to be accessed and analyzed.

To facilitate the discovery of putative cancer immunotherapies and their respective targets, several types of databases and tools have been developed for predicting or detecting tumor-specific neoantigen and immune antigen,[6–13] compiling cancer immune checkpoints and their modulators [14], evaluating genetic variants on tumor immune infiltration [15], exploring

CONTACT Godfrey Chi-Fung Chan gcfchan@hku.hk; Wenliang Zhang zhangwl2@hku-szh.org Department of Pediatrics, The University of Hong Kong-Shenzhen Hospital, Shenzhen, 518058, China; Weizhong Li liweizhong@mail.sysu.edu.cn Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, 510080 China; Dongsheng Yu yudsh@mail.sysu.edu.cn Hospital of Stomatology, Guangdong Provincial Key Laboratory of Stomatology, Guanghua School of Stomatology, Sun Yat-sen University, Guangzhou, 510055, China.

*Those authors contribute equally as co-first author.
#These are co-correspondence authors.

Supplemental data for this article can be accessed on the publisher's website

molecular mechanism of traditional Chinese medicine on cancer immunology [16], and providing oncolytic virus-based cancer immunotherapy.[17,18] For example, TSNAdb [11] is a database for tumor-specific neoantigens from immunogenomics data analysis; pVAC-Seq [12] is a genome-guided tool for identifying tumor neoantigens; pTuneos [13] is another tool for prioritizing tumor neoantigens from next-generation sequencing data; and CancerImmunityQTL [15] is a database to systematically evaluate the impact of genetic variants on tumor immune infiltration. In October 2020, Zhang et al. developed the CKTTD [14] database via enhanced text-mining system with manual curation. CKTTD provides the association data between cancer immune checkpoint therapies and their targets. These databases and tools have facilitated the discovery of putative immunotherapies and targets for human cancers. However, the lack of public accessible common database for in-depth validation of cancer immunotherapy associated with their predictive biomarkers, targets, control therapies, clinical efficacy, and adverse events

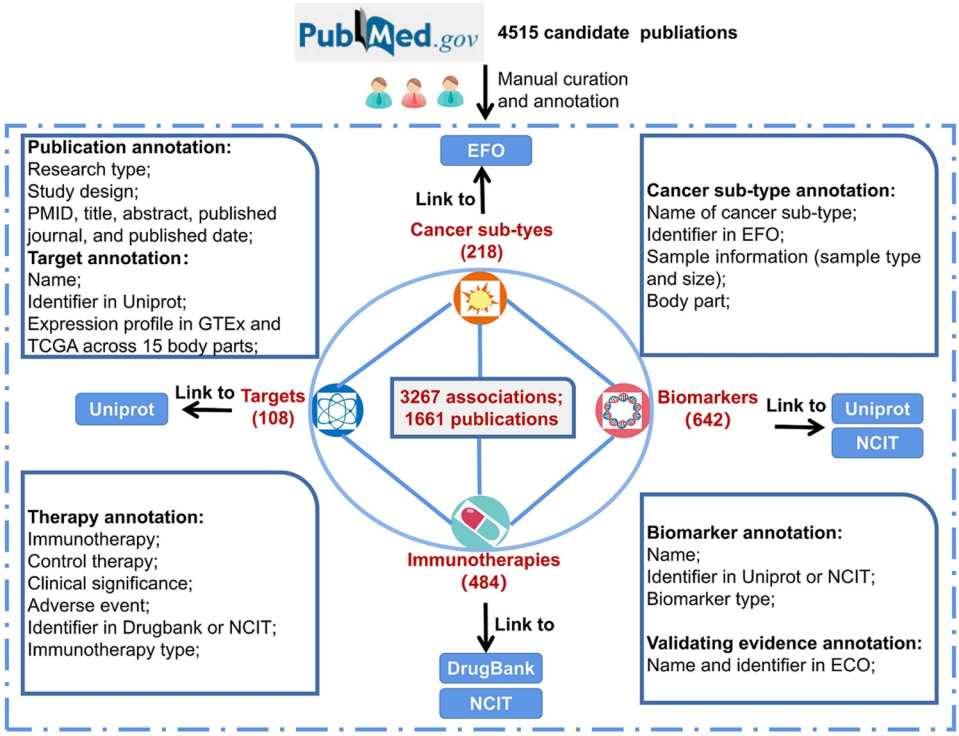


Figure 1. The data curation and annotation framework of canimmunother.

Table 1. Database contents and features of CanImmunother compared with CKTTDB

Content and web applications	CanImmunother	CKTTDB	CanImmunother/CKTTDB (Fold change)
Non-redundant associations	2646	210	**12.60**
Cancer sub-types	218	33	**6.61**
Immunotherapy	484 (Various types of immunotherapy *)	53 (Immune checkpoint therapy only)	**9.13**
Control therapy	121	None	–
Biomarker	642	None	–
Target	108	105	1.03
Clinical efficacy	Yes	None	–
Adverse event	Yes	None	–
Sample information (type and size)	Yes	None	–
Study design	Yes	None	–
Research type	Yes	None	–
Gene expression profile for target	Yes	Yes	–
Association score	Yes	Yes	–
Analysis application	Association prioritization; Network visualization	None	**More web applications**
Data quality	Manual curation on literature	Curation with an enhanced text-mining system and data integration	**Higher quality**

Note: * The various types of immunotherapy include immune checkpoint therapy, tumor vaccine, immune-related cytokine, cellular immunotherapy, oncolytic viruses, and their combination with other non-immunotherapy, such as chemotherapy, target therapy, radiotherapy, surgery, chemoradiotherapy, and hormone therapy (Figure 2c).

remains the current bottleneck. The establishment of such database can help clinicians and researchers to identify and develop novel immunotherapy options for cancer patients.

To tackle these problems, we developed the CanImmunother database (http://www.biomedical-web.com/cancerit/) through manual curation on 4515 publications. CanImmunother is the first comprehensive database to provide experimentally validated cancer immunotherapies associated with their biomarkers, targets, and control therapies, enabling the comparison and clarification of their respective clinical efficacy and the adverse events. The association data in CanImmunother were consistently annotated with standard terminology and ontology (Figure 1). Currently, CanImmunother provides 3267 experimentally validated associations between 218 cancer sub-types and 484 immunotherapies with 642 biomarkers, 108 targets, and 121 control therapies (Table 1). Each association was reviewed, manually curated by multiple professional curators and incorporated with valuable annotation and cross-references. CanImmunother offers user-friendly web interfaces and web applications such as excel table, association prioritization, and network visualization, to help clinicians and researchers in identifying and discovering advanced cancer immunotherapies and their respective biomarkers and targets. CanImmunother will be able to serve as a useful resource to improve our insight and to facilitate the identification and discovery of advanced immunotherapy options for patients with cancer.

Materials and methods

Literature reviewing and manual curation

To collect literature data manually, according to our previous method [19], we searched the National Center for Biotechnology Information (NCBI) PubMed database [20] for candidate publications that described the studies for human cancer immunotherapy. Search terms and their combinations used in the search strategy included cancer, carcinoma, neoplasm, tumor, leukemia, lymphoma, melanoma, malignancy, immunotherapy, immune checkpoint, and specific immune checkpoint agent and nonimmune checkpoint agent names, which were described in the systematic review publications.[2,21] 4515 candidate publications were retrieved before August 2020. We then filtered the abstracts of these candidate publications based on two criteria. First, the publications are original research literature (including clinical and basic research papers, as well as case reports). However, review and commentary papers were excluded. Second, the publications reported experimentally validated human cancer immunotherapy, such as immune checkpoint therapy, tumor vaccine, immune related cytokine, cellular immunotherapy, and so on. 1932 publications were retained after filtering abstracts. Furthermore, the full texts of 1932 publications were reviewed and manually curated by multiple professional curators to collect and annotate the association data for cancer immunotherapy. All the curators are professionals in tumor immunology and genetics.

To ensure the quality in the data curating process based on our previous manual curation method [19], we randomly selected 10.35% (200/1932) of the publications initially. They were curated and discussed among all curators to achieve a consensus for data manual curation. Second, each publication was reviewed and manually curated by at least two curators. Third, if the curated data from the same publication by different curators were not consistent, a third curator would review the publication and further discuss with the team to reach a consistent decision. The collected data include cancer subtype, immunotherapy and its control therapy, biomarker, target, adverse event, sample information (patients/cell lines/animal models and size), PubMed identifier (PMID), research type, study design, and validation evidence, from the full text of the supporting publications (Figure 1). The study design information collected from publications, such as clinical trials of different phases, meta-analysis, and retrospective cohort study, enables users to clearly evaluate their reliability and methodological bias. In order to make the association data more useful, each association was annotated with a clinical significance, which is a brief summing-up description annotated by professional curators based on the information from the original publications (Figure 1).

Data annotation

To make the extracted data consistent and accessible, the cancer sub-types, immunotherapies and their control therapies, biomarkers, targets, and validating evidences in CanImmunother were manually annotated with standard terminology and ontology (Figure 1). Cancer sub-types were annotated by Experimental Factor Ontology (EFO), which provides a systematic description of many experimental variables available in the European Bioinformatics Institute databases and for many international projects [22]. The majority of immunotherapies and their control therapies were annotated by DrugBank [23], while the rest of the therapies, of which DrugBank does not cover, were annotated by National Cancer Institute Thesaurus (NCIT) OBO Edition [24]. The UniProtKB [25] and NCIT OBO Edition [24] resources were adopted to annotate biomarkers and targets. The Evidence & Conclusion Ontology (ECO) [26] was used to annotate validation evidences. The multiple terminology and ontology resources were integrated for the data annotation that is to guarantee the accuracy. Moreover, the targets in CanImmunother were annotated with the expression (mean of FPKM) profile on gene level across 15 body parts in the Genotype-Tissue Expression (GTEx) [27] and The Cancer Genome Atlas (TCGA) [28] resources by using the PreMedKB tool [29]. In addition, each association was also systematically annotated with title, abstract, published journal, and published date of the supporting publication. The information of resources used in CanImmunother is detailed in Supplemental Table 1.

Association score

In exploiting large collections of aggregated association data, one of the main problems is how to prioritize and interpret the association data [30]. According to our previous methods,[19,31,32] we refined a scoring model to compute an association score for each association data with or without predictive biomarker in

CanImmunother, respectively. It was based on two evidential metrics: the research types of the supporting publications (e.g., clinical research, basic research, and case report) and the number of supporting publications. Different research types were assigned with different weights based on their reliability. Moreover, as a larger number of publications can enhance association score for the same association, a harmonic sum function [19,31,32] was employed not only to compute the score of the same association with multiple supporting publications but also to dampen the effect of data volume or quantity. Finally, association scores were normalized to limit their range from 0 to 1.0 for better interpretation. The computation steps of association score were described in detail at the "Help" webpage of the database (http://www.biomedical-web.com/cancerit/help.jsp).

Web implementation

The web database was built with Spring MVC and jQuery AJAX frameworks. All association data in CanImmunother was organized in MySQL. The data accessing and processing programs were written in Java. The web interface was implemented by JavaScript, HTML5, and CSS3. The vis.js widget (http://www.visjs.org) was implemented to display the networks on the web-pages. The widgets of excel-bootstrap-table-filter-bundle.js and dataTables.bootstrap.js were used to implement the filter tables on the web pages. CanImmunother is freely available at the website http://www.biomedical-web.com/cancerit/.

Results

Data contents

Currently, CanImmunother provides 3267 experimentally validated associations between 218 cancer sub-types across 34 body parts and 484 immunotherapies with 642 biomarkers, 108 targets, and 121 control therapies across 1661 supporting publications (Figure 1 & Table 1). Each association was incorporated with valuable annotation and cross-references and given with a unique accession number (e.g., CANIT0000001). The top six cancer sub-types with the largest numbers of associations are melanoma, non-small cell lung carcinoma, renal cell carcinoma, colorectal cancer, urothelial carcinoma, and prostate cancer (Figure 2a). The top six body parts with the largest numbers of associations are skin, lung, kidney, blood and lymph node, bladder, and intestines (Figure 2b). CanImmunother provides various types of immunotherapy, which include immune checkpoint therapy, tumor vaccine, immune-related cytokine, cellular immunotherapy, oncolytic virus, and their combination, or they combined with other non-immunotherapy, such as chemotherapy, target therapy, radiotherapy, hormone therapy, chemoradiotherapy, and surgery (Figure 2c). The top six largest numbers of associations in CanImmunother are related to immune checkpoint therapy, immune checkpoint therapy plus chemotherapy, tumor vaccine, immune checkpoint therapy plus target therapy, immune checkpoint therapy plus radiotherapy, and immune checkpoint therapy plus tumor vaccine (Figure 2c). Importantly, CanImmunother also provides 642 biomarkers and 108 targets for the cancer immunotherapy associations. The top six number of biomarker types are gene expression signature on/in tumor cell, mutational status, disease status, gene expression signature on/in immune cell, gene expression signature in serum, and cell percentage/counts in blood (Figure 2d). Remarkably, the biomarker of *PD-L1* expression level makes up the most common associations and is associated with 92 cancer sub-types and 36 immunotherapies to form 223 associations in CanImmunother. In addition, approximately 74.47% (2433/3267), 3.86% (126/3267), and 21.67% (708/3267) of the associations are supported with clinical research publications, basic research publications, and case reports, respectively (Figure 2e).

Browse and search

CanImmunother (http://www.biomedical-web.com/cancerit/) provides user-friendly web interfaces and web services to enable users to search, browse, and analyze the association data, as well as to download and submit new associations for further investigation and integration. The "Browse all" webpage presents all of the associations in a table to allow users to filter for the interesting associations through the "search" box. Moreover, we implemented "Browse" sub-web-pages to allow users to browse interesting associations through different body parts, cancer sub-types, biomarker types, and immunotherapy types. In addition, at the "Home" web-page, the symbols in the word-cloud diagrams can be navigated to browse their entry details.

To enable users to quickly retrieve the interesting associations and their supporting publications, CanImmunother provides a search application using cancer sub-type, immunotherapy, biomarker, or target with setting filtration parameters, such as body part, immunotherapy type, biomarker type, and research type (Figure 3a). The search application also provides a smart assistance by listing the closest entries to that expectation. The resulting associations are shown in a brief table that displays key information, including cancer sub-types, immunotherapies and their control therapies, biomarkers, targets, study designs, research types, and PMIDs (Figure 3b). Moreover, the "Detail" button in the result table links to further webpages for extra information of the association (Figure 3b). The extra information includes clinical significance and adverse event of the immunotherapy compared with the control therapy, sample information, validating evidence, and other information of the supporting publication (Figure 3b). In addition, CanImmunother provides external links to the related reference resources, such as NCBI PubMed, EFO, DrugBank, NCIT OBO Edition, UniProtKB, and ECO (Figure 3b). Furthermore, for the immunotherapy associated target, CanImmunother provides its expression (mean of FPKM) profile on gene level in the GTEx and TCGA resources across 15 body parts (Figure 3c).

Case study 1: assisting identification of better cancer immunotherapies and their predictive biomarkers

A simple search in CanImmunother with the keyword "non-small cell lung carcinoma" can obtain 624 associations between non-small cell lung carcinoma (NSCLC) and 97

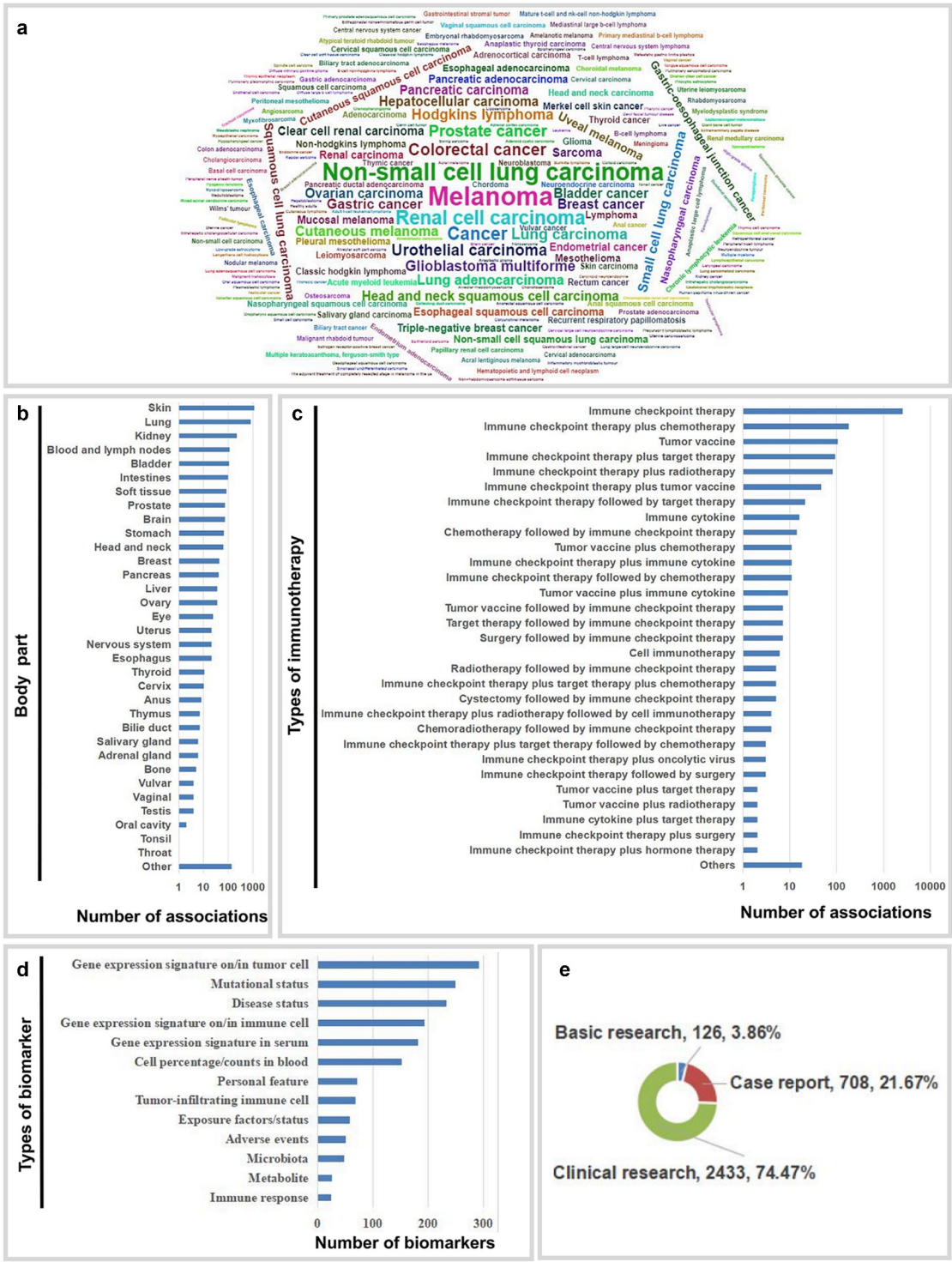


Figure 2. The landscape of association data in canimmunother. (a) A word-cloud diagram shows the association landscape of 218 cancer sub-types in canimmunother. Larger sizes and more central locations of the cancer sub-type symbols in the diagram indicate more association data in the database. (b) The numbers of associations in different types of body part. (c) The number of associations in different types of immunotherapy. Others include cell immunotherapy followed by immune checkpoint therapy, cell immunotherapy followed by target therapy, cell immunotherapy plus immune cytokine, immune checkpoint therapy followed by immune checkpoint therapy plus target therapy, immune checkpoint therapy followed by radiotherapy, immune checkpoint therapy plus anti-angiogenesis therapy, immune checkpoint therapy plus cell immunotherapy, immune checkpoint therapy plus cell immunotherapy plus radiotherapy, immune checkpoint therapy plus chemotherapy followed by surgery, immune checkpoint therapy plus chemotherapy or radiotherapy, immune checkpoint therapy plus oncolytic virus plus chemotherapy, immune checkpoint therapy plus systemic therapy, immune cytokine followed immune checkpoint therapy, immune cytokine plus chemotherapy, immune cytokine plus target therapy followed by immune checkpoint therapy, radiotherapy followed by immune checkpoint therapy plus chemotherapy, surgery plus target therapy followed by immune checkpoint therapy, target therapy followed by immune checkpoint therapy plus hormone therapy. (d) The number of immunotherapy related biomarkers for different types of biomarker. (e) The percentage and number of associations in different research types of the supporting publication.

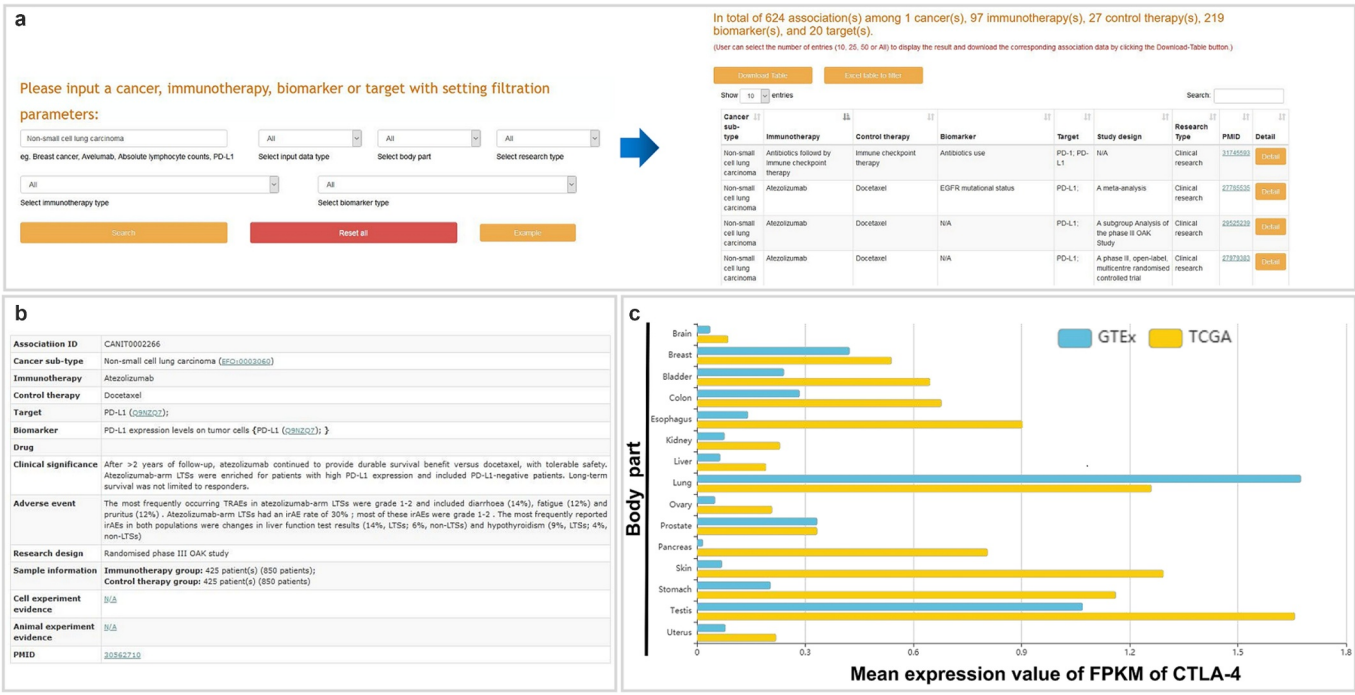
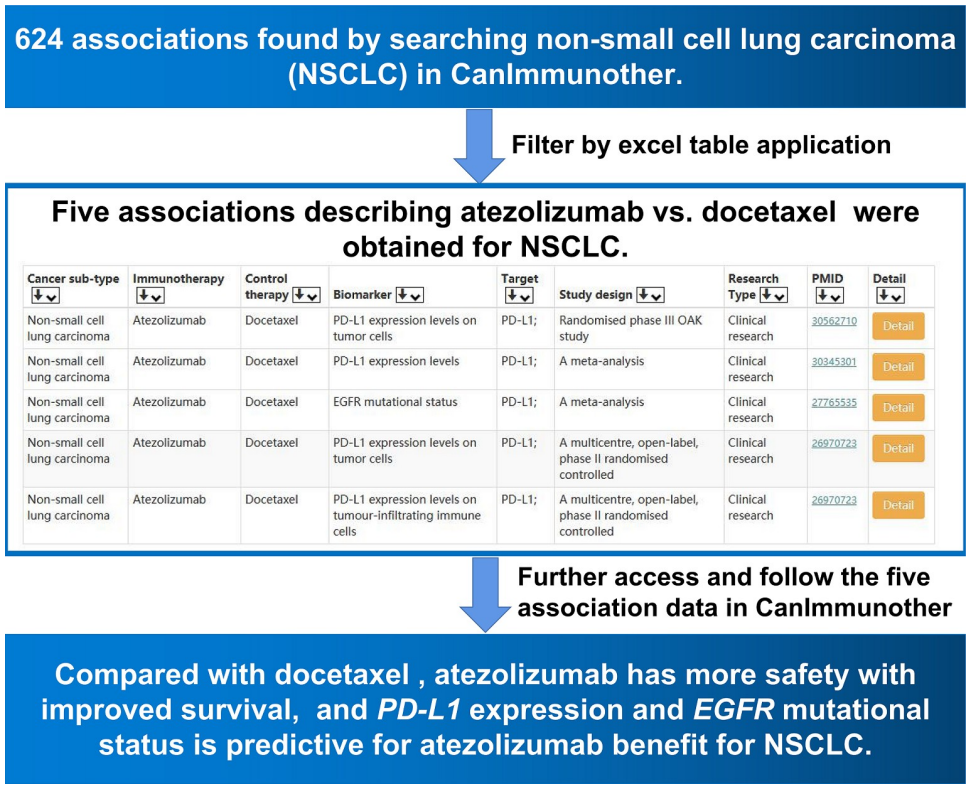


Figure 3. The web interface of search and excel table application. (a) A resulting table by searching words like "non-small cell lung carcinoma" indicates non-small cell lung carcinoma associating with 97 immunotherapies, 219 biomarkers, 20 targets, and 27 control therapies to form 624 associations. (b) Each association in CanImmunother was manually curated and annotated with valuable annotation. (c) The mean expression profile of fragments per kilobase of exon model per million mapped fragments (FPKM) of *CTLA-4* in the GTEx and TCGA resources across 15 body parts.



Cancer sub-type	Immunotherapy	Control therapy	Biomarker	Target	Study design	Research Type	PMID	Detail
Non-small cell lung carcinoma	Atezolizumab	Docetaxel	PD-L1 expression levels on tumor cells	PD-L1;	Randomised phase III OAK study	Clinical research	30562710	Detail
Non-small cell lung carcinoma	Atezolizumab	Docetaxel	PD-L1 expression levels	PD-L1;	A meta-analysis	Clinical research	30345301	Detail
Non-small cell lung carcinoma	Atezolizumab	Docetaxel	EGFR mutational status	PD-L1;	A meta-analysis	Clinical research	27765535	Detail
Non-small cell lung carcinoma	Atezolizumab	Docetaxel	PD-L1 expression levels on tumor cells	PD-L1;	A multicentre, open-label, phase II randomised controlled	Clinical research	26970723	Detail
Non-small cell lung carcinoma	Atezolizumab	Docetaxel	PD-L1 expression levels on tumour-infiltrating immune cells	PD-L1;	A multicentre, open-label, phase II randomised controlled	Clinical research	26970723	Detail

Figure 4. CanImmunother assists users to identify better cancer immunotherapies and their predictive biomarkers. By using the excel table application in CanImmunother, five associations, which describe atezolizumab versus docetaxel with their predictive biomarkers and targets for NSCLC, were efficiently obtained from 624 associations. By further accessing and following the five associations in CanImmunother, we identify that atezolizumab immunotherapy is a better therapy for advanced NSCLC in comparison with docetaxel chemotherapy, and PD-L1 expression and EGFR mutational status are predictive for atezolizumab benefit.

immunotherapies with 219 biomarkers, 20 targets, and 27 control therapies (Figure 3a). In order to allow users to filter out interesting associations efficiently from those large numbers of associations, an excel table application was implemented for combinational filtration. For example, to compare the clinical efficacy and adverse event of atezolizumab immunotherapy with docetaxel chemotherapy in NSCLC, we used the excel table application in CanImmunother and efficiently obtained five associations, which describe atezolizumab versus docetaxel with their biomarkers and targets for NSCLC, from 624 associations (Figure 3a & 4). Moreover, we further accessed and followed the five associations in CanImmunother. The five associations suggested that (1) compared with docetaxel, atezolizumab is safer with better survival in patients with advanced NSCLC, regardless of *PD-L1* expression; however, higher *PD-L1* levels on tumor cells and tumor-infiltrating immune cells were likely to correlate with better outcome; (2) compared with docetaxel, atezolizumab significantly improved survival in patients with advanced NSCLC in overall and in the *EGFR* wild-type subgroup, but not in the *EGFR* mutant subgroup. So, through searching, filtering, accessing, and following the interesting association data in CanImmunother, we identified that atezolizumab immunotherapy is a better therapy for advanced NSCLC in comparison with docetaxel chemotherapy, and *PD-L1* expression and *EGFR* mutational status are predictive biomarkers for the beneficial effect of atezolizumab.

Case study 2: association prioritization to prioritize cancer immunotherapies and their predictive biomarkers

Multiple publications probably support the same association. To enable further analysis of the associations, 2646 non-redundant associations were produced from the 3267 associations in CanImmunother by removing data redundancy based on the supporting publications. Each non-redundant association was assigned with an association score for further prioritization and analysis. The assignment of association score was described in detail at the "Materials and Methods" and the "Help" webpage. To promptly evaluate the associations, an association prioritization application was established based on the non-redundant association data in CanImmunother. Association prioritization application allows users to retrieve a cancer sub-type, an immunotherapy, a biomarker, and a target alone or together for prioritizing cancer immunotherapy association data with their predictive biomarkers and targets. For instance, we searched by a keyword like "serum lactate dehydrogenase level" and promptly identified that seven cancer sub-types and ten immunotherapies are associated with the biomarker of serum lactate dehydrogenase level to form sixteen associations. The top three associated immunotherapies are ipilimumab, nivolumab, and pembrolizumab, and the top three associated cancer sub-types are melanoma, cutaneous melanoma, and non-small cell lung carcinoma (Figure 5a). The results allow sorting by association scores and filtering by specific cancer sub-types, immunotherapies, biomarkers, and targets through the "search" box (Figure 5a). In addition, each prioritized association data can be optionally visualized in an interactive network diagram to display their relationships (Figure 5b). Figure 5b displays all experimentally validated biomarkers for uveal melanoma with nivolumab therapy to target PD-1 protein.

Case study 3: network visualization to discover potential cancer immunotherapies and their predictive biomarkers and targets

To explore the relationships of the experimentally validated association data in CanImmunother for discovering cancer immunotherapies and their predictive biomarkers and targets, a network visualization application was implemented in CanImmunother. The application allows users to input a set of cancer sub-types, immunotherapies, biomarkers, and targets and to construct interactive networks to display their relationships. For instance, we entered an input of "head and neck squamous cell carcinoma and nasopharyngeal squamous cell carcinoma" and constructed an interactive network to explore the relationships of the two cancer sub-types with immunotherapies, biomarkers, and targets (Figure 6). The interaction diagram shown that head and neck squamous cell carcinoma has nine predictive biomarkers such as *PD-L1* and *PD-L2* expression level, HPV infection, tumor mutational burden, and DNA mismatch repair-deficient or microsatellite instability for nivolumab, pembrolizumab, and durvalumab therapy to, respectively, target PD-1 and PD-L1 proteins, while nasopharyngeal squamous cell carcinoma has only one predictive biomarker of EBV-positive for MVA-EL vaccine therapy to target EBNA1 and LMB2 proteins (Figure 6). Moreover, the interaction diagram also implied that the nine biomarkers of head and neck squamous cell carcinoma may be potential predictive biomarkers for nasopharyngeal squamous cell carcinoma with nivolumab, pembrolizumab, and durvalumab therapy, but need to further validate and confirm by clinical trials (Figure 6). Furthermore, the network diagram indicates that for patients with head and neck squamous cell carcinoma, *PD-L1* expression level on tumor cells is the common biomarker of pembrolizumab and durvalumab by targeting PD-1 and PD-L1 proteins, respectively (Figure 6). So, the network visualization application not only demonstrates relationships between different cancer sub-types and immunotherapies with their respective biomarkers and targets but also discovers and develops potential cancer immunotherapies and their predictive biomarkers and targets from the experimentally validated association data in CanImmunother. As the interactive networks may consist of many nodes and even more interactions (Figure 5b and Figure 6), CanImmunother offers users filtering function to hide those interactions that are less interesting (Figure 6). When selecting or unselecting some of the nodes, the interactive network will be changed accordingly, and a sub-network of the whole interactive network will then be displayed. In addition, all nodes in the network allow adjustment except for node legends.

Data access

Web service application programming interfaces (APIs) were implemented for programmatic access of association data in the CanImmunother database. The accessing data by the APIs are available in the universal JSON formats. Moreover, all association data in the database can be freely downloaded for

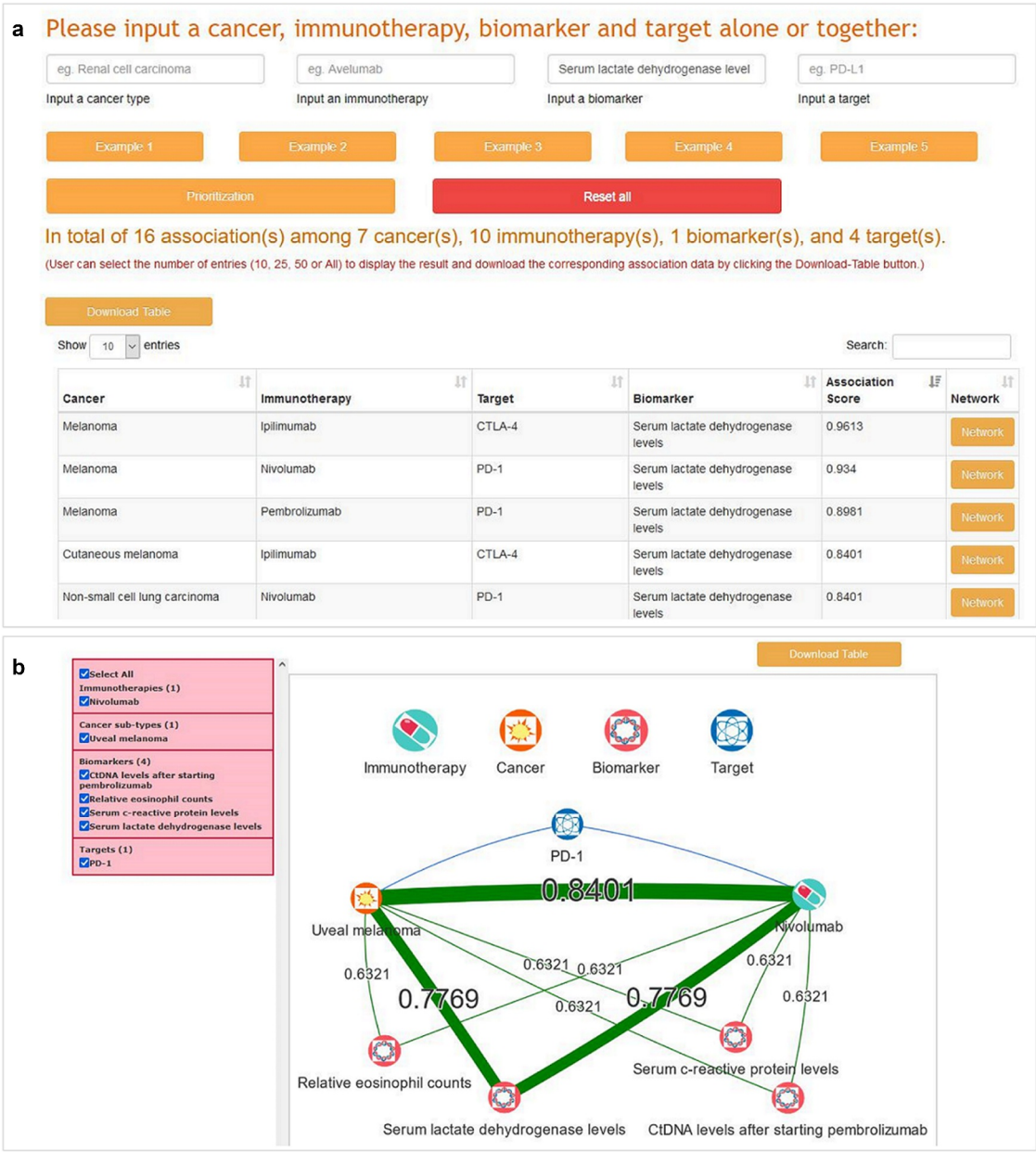


Figure 5. The web interface of association prioritization and network visualization in canimmunother. (a) A resulting table prioritizes cancer sub-types, immunotherapies, and targets associated with the biomarker of serum lactate dehydrogenase level. (b) A network diagram displays all experimentally validated biomarkers for uveal melanoma with nivolumab therapy to target PD-1 protein. Green lines connect biomarker with cancer sub-type and immunotherapy, while blue lines connect target with cancer sub-type and immunotherapy. The values on the green lines are association scores. The thicker green lines represent larger association scores, and the thinner green lines for smaller association scores. The association scores are ranging from 0 to 1.0.

further investigation and integration. In addition, we encourage users to submit their new experimentally validated association data for cancer immunotherapy. Once checked and approved by our submission review committee, the submission data will be included in a future release. Finally, a detailed tutorial for the database is available on the 'Help' web-page.

Discussion and conclusion

As the publications for cancer immunotherapy exploded exponentially in recent years, a large amount of experimentally validated association data for cancer immunotherapy remains hidden in the literature and is difficult to access and utilize. Therefore, the curation and analysis of these association data from publications can economically utilize the existing data fully for expediting translational research and application of cancer immunotherapy. In this study, we designed and constructed a comprehensive database called CanImmunother through manually curating cancer immunotherapy association data from peer-review publications. We consistently correlated these association data with valuable annotation. As far as we are aware, CanImmunother is the first comprehensive and high-quality database to provide experimentally validated information of cancer immunotherapy in association with their biomarkers, targets, and control therapies. This database

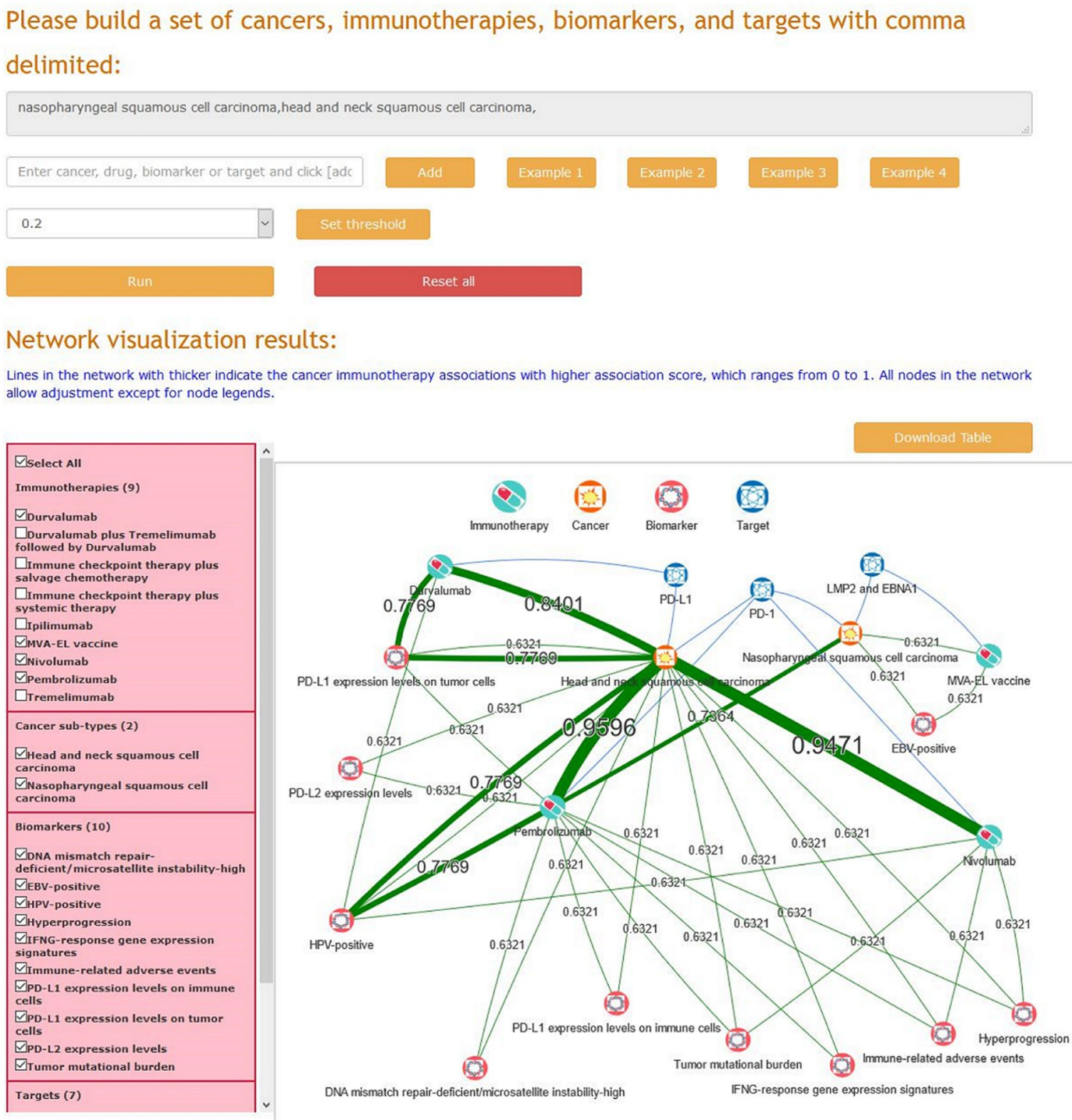


Figure 6. Network visualization explores relationships of the experimental association data in canimmunother to discover potential cancer immunotherapies and their predictive biomarkers and targets. The network diagram displays the relationships of head and neck squamous cell carcinoma and nasopharyngeal squamous cell carcinoma with their immunotherapies, biomarkers, and targets. Green lines connect biomarker with cancer sub-type and immunotherapy, while blue lines connect target with cancer sub-type and immunotherapy. The values on the green lines are association scores. The thicker green lines represent larger association scores, and the thinner green lines for smaller association scores. The association scores are ranging from 0 to 1.0.

can help to compare and clarify on clinical efficacy and adverse events of specific immunotherapy on a particular cancer type.

To accelerate the progress of cancer immunotherapy, several useful computational resources for cancer immunology have been developed recently.[7,14–17] Different from these computational resources such as CKTTD, our CanImmunother database focuses on providing various cancer immunotherapy association data with experimentally validated to help clinicians and researchers for identification and discovery of advanced immunotherapy options for patients with cancer. In Case study 1, we used CanImmunother to access and follow the association data about atezolizumab immunotherapy versus docetaxel chemotherapy in advanced NSCLC and identified that atezolizumab is a better option in terms of efficacy and safety when compared with docetaxel, and *PD-L1* expression and *EGFR* mutational status are predictive biomarkers for atezolizumab benefit. Moreover, in Case study 3, our network visualization application improved our insight on potential cancer immunotherapy and their predictive biomarkers and targets through exploring the relationships between different cancer sub-types, immunotherapies and their biomarkers and targets.

Compared with CKTTD, our CanImmunother database significantly outperforms CKTTD in data coverage, data quality, and application feature (Table 1). First, our CanImmunother database covers various types of cancer immunotherapy, including immune checkpoint therapy, tumor vaccine, immune related cytokine, cellular immunotherapy, oncolytic virus, and their combination or they combined with other non-immunotherapy, such as chemotherapy, target therapy, radiotherapy, chemoradiotherapy and hormone therapy, while

CKTTD provides information related to single agent of immune checkpoint therapy only (Figure 2 c & Table 1). Second, each association in CanImmunother was annotated with extra valuable annotation, including predictive biomarkers, control therapies, clinical efficacy and adverse events, sample information, and the research type and study design of the supporting publications, while associations in CKTTD do not have similar function (Figure 3 c & Table 1). Third, CanImmunother offers extra applications for further analysis on the association data, including excel table application, association prioritization, and network visualization (Table 1). Finally, the number of associations, cancer sub-types, and immunotherapies contained in CanImmunother are approximately 12.60-, 6.61-, and 9.13-fold of those in CKTTD (Table 1). In addition, each association in CanImmunother was designated with an association score for prioritization by assigning different weights to different research types of the supporting publications. The comparison of data contents and features between CanImmunother and CKTTD is shown in Table 1.

With the rapid advancement of cancer immunotherapy, more and more experimentally validated cancer immunotherapy data are expected to be reported in the near future. To serve the research communities in fully utilizing the vast amount of data, we will update CanImmunother every six months and constantly improve it with more features and functionalities. Currently, all data in CanImmunother were manually curated from peer-review publications, thus without association data from the international collaboration projects and resources, such as TCGA [28], International Cancer Genome Consortium (ICGC) [33], and Gene Expression Omnibus (GEO) [34]. Therefore, we plan to enrich new association data through analyzing cancer immunotherapy-related datasets in those projects and resources. In addition, we will also develop and integrate more computational resources [35–37] and tools [8,9,13] in the database to annotate and analyze those association data, such as similarity prediction between new chemicals and known cancer immunotherapeutic agents. In conclusion, as a timely and helpful resource, CanImmunother will enhance our insight on cancer immunotherapy, and to facilitate the identification and discovery of advanced immunotherapy options for patients with cancer.

Abbreviations

APIs - Application Programming Interfaces
CTLA-4 - Cytotoxic T-lymphocyte-associated protein 4
EBV - Epstein Barr virus
ECO - Evidence & Conclusion Ontology
EFO - Experimental Factor Ontology
FPKM - Fragments per kilobase of exon model per million mapped fragments
GEO - Gene Expression Omnibus
GTEx - Genotype-Tissue Expression
HPV - Human papilloma virus
ICGC - International Cancer Genome Consortium
NCBI - National Center for Biotechnology Information
NSCLC - Non-small cell lung carcinoma
NCIT - National Cancer Institute Thesaurus
PD-1 - Programmed cell death protein 1
PD-L1 - Programmed cell death-ligand 1
PMIDs - PubMed identifiers
TCGA - The Cancer Genome Atlas

Acknowledgments

We thank the support from The Clinical, Translational and Basic Research Laboratory of The University of Hong Kong - Shenzhen Hospital and Sun Yat-sen University.

Disclosure statement

The authors declare that they have no competing interests.

Funding

This work was supported by Guangdong Basic and Applied Basic Research Foundation, China [grant number 2020A1515110528] and China Postdoctoral Science Foundation [grant number 2021M693302] to Wenliang Zhang; The Strategic Priority CAS Project [grant number XDB38000000], the National Key R&D Program of China [grant number 2018YFB0204403], and the Shenzhen Basic Research Fund [grant number RCYX20200714114734l9 & JCYJ20170413093358429] to Yanjie Wei; The National Key R&D Program of China [grant number 2018YFC0910401 & 2016YFC0901604], the Natural Science Foundation of Guangdong Province [grant number 2021A1515012108], the Guangdong Project [grant number 2017GC010608], and the Support Scheme of Guangzhou for Leading Talents in Innovation and Entrepreneurship [grant number 2020007] to Weizhong Li; The National Natural Science Foundation of China [grant number 81873711] to Dongsheng Yu; The Fundamental Research Funds of the Central Universities, Sun Yat-sen University [grant number 19ykpy86] and China Postdoctoral Science Foundation [grant number 2020M673023] to Binghui Zeng; The Sanming Project of Medicine (Shenzhen) [grant number SZSM201911016]

Notes on contributors

Wenliang Zhang: Project administration, Conceptualization, Methodology, Web implementation, Data curation and validation, Funding acquisition, and Writing—Original draft, Review & Editing.

Binghui Zeng: Conceptualization, Resources, Methodology, Data curation and validation, Funding acquisition, and Writing—Review & Editing.

Huancai Lin: Data curation and validation, and Writing—Review & Editing.

Wen Guan: Data curation and validation.

Mo Jing: Web implementation.

Song Wu: Data validation.

YanJie Wei: Resources and Writing—Review & Editing

Qianshen Zhang: Resources, Data curation, and Writing—Review & Editing.

Dongsheng Yu: Resources, Methodology, Data curation and validation, and Writing—Review & Editing.

Weizhong Li: Resources, Funding acquisition, and Writing—Review & Editing.

Godfrey Chi-Fung Chan: Resources, Funding acquisition, and Writing—Review & Editing

ORCID

Wenliang Zhang http://orcid.org/0000-0003-0454-6935
Dongsheng Yu http://orcid.org/0000-0002-2176-9308
Weizhong Li http://orcid.org/0000-0002-9003-7733
Godfrey Chi-Fung Chan http://orcid.org/0000-0001-5032-8985

Reference

1. Hegde PS, Chen DS. Top 10 challenges in cancer immunotherapy. Immunity. 2020;52(1):17–35. doi:10.1016/j.immuni.2019.12.011.
2. Marin-Acevedo JA, Soyano AE, Dholaria B, Knutson KL, Lou Y. Cancer immunotherapy beyond immune checkpoint inhibitors. J Hematol Oncol. 2018;11(1):8. doi:10.1186/s13045-017-0552-6.
3. Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. Science. 2018;359(6382):1350. doi:10.1126/science.aar4060.
4. Galluzzi L, Chan TA, Kroemer G, Wolchok JD, Lopez-Soto A. The hallmarks of successful anticancer immunotherapy. Sci Transl Med. 2018;10(459):459. doi:10.1126/scitranslmed.aat7807.
5. Pauken KE, Dougan M, Rose NR, Lichtman AH, Sharpe AH. Adverse events following cancer immunotherapy: obstacles and opportunities. Trends Immunol. 2019;40(6):511–523. doi:10.1016/j.it.2019.04.002.
6. Nakamura Y, Komiyama T, Furue M, Gojobori T, Akiyama YCIG. DB: the database for human or mouse immunoglobulin and T cell receptor genes available for cancer studies. BMC Bioinform. 2010;11(1):398. doi:10.1186/1471-2105-11-398.
7. Tan X, Li D, Huang P, Jian X, Wan H, Wang G, Li Y, Ouyang J, Lin Y, Xie L. dbPepNeo: a manually curated database for human tumor neoantigen peptides. Database (Oxford). 2020;2020:2020. doi:10.1093/database/baaa004.
8. Jaravine V, Mosch A, Raffegerst S, Schendel DJ, Frishman D. Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues. BMC Cancer. 2017;17(1):892. doi:10.1186/s12885-017-3854-8.
9. Wu J, Wang W, Zhang J, Zhou B, Zhao W, Su Z, Gu X, Wu J, Zhou Z, Chen S. DeepHLApan: approach for neoantigen prediction considering both HLA-Peptide binding and immunogenicity. Front Immunol. 2019;11:2559. doi:10.3389/fimmu.2019.02559.
10. Zhou Z, Lyu X, Wu J, Yang X, Wu S, Zhou J, Gu X, Su Z, Chen S. TSNAD: an integrated software for cancer somatic mutation and tumour-specific neoantigen detection. R Soc Open Sci. 2017;4(4):170050. doi:10.1098/rsos.170050.
11. Wu J, Zhao W, Zhou B, Su Z, Gu X, Zhou Z, Chen S. TSNAdb: a database for tumor-specific neoantigens from immunogenomics data analysis. Genomics Proteomics Bioinformatics. 2018;16(4):276–282. doi:10.1016/j.gpb.2018.06.003.
12. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, Griffith M. pVAC-Seq: a genome-guided in silico approach to identifying tumor neoantigens. Genome Med. 2016;8(1):11. doi:10.1186/s13073-016-0264-5.
13. Zhou C, Wei Z, Zhang Z, Zhang B, Zhu C, Chen K, Chuai G, Qu S, Xie L, Gao Y, et al. pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. Genome Med. 2019;11(1):67. doi:10.1186/s13073-019-0679-x.
14. Zhang Y, Yao Y, Chen P, Liu Y, Zhang H, Liu H, Liu Y, Xu H, Tian X, Wang Z, et al.. Checkpoint therapeutic target database (CKTTD): the first comprehensive database for checkpoint targets and their modulators in cancer immunotherapy. J Immunother Cancer. 2020;8(2):2. doi:10.1136/jitc-2020-001247.
15. Tian J, Cai Y, Li Y, Lu Z, Huang J, Deng Y, Yang N, Wang X, Ying P, Zhang S, et al.. CancerImmunityQTL: a database to systematically evaluate the impact of genetic variants on immune infiltration in human cancer. Nucleic Acids Res. 2020. doi:10.1093/nar/gkaa805.
16. Liu Z, Cai C, Du J, Liu B, Cui L, Fan X, Wu Q, Fang J, Xie L. TCMIO: A. Comprehensive Database of traditional Chinese medicine on immuno-oncology. Front Pharmacol. 2020;11:439. doi:10.3389/fphar.2020.00439.
17. Lathwal A, Kumar R, OvirusTdb RG. A database of oncolytic viruses for the advancement of therapeutics in cancer. Virology 2020;548:109–116. doi:10.1016/j.virol.2020.05.016.
18. Lathwal A, Kumar R, Raghava G. Computer-aided designing of oncolytic viruses for overcoming translational challenges of cancer immunotherapy. Drug Discov Today. 2020;25(7):1198–1205. doi:10.1016/j.drudis.2020.04.008.
19. Zhang W, Zeng B, Yang M, Yang H, Wang J, Deng Y, Zhang H, Yao G, Wu S, Li W. ncRNAVar: a manually curated database for identification of noncoding RNA variants associated with human diseases. J Mol Biol. 2021;433(11):166727. doi:10.1016/j.jmb.2020.166727
20. National Genomics Data Center Members and Partners. Database resources of the national genomics data center in 2020. Nucleic Acids Res. 2020;48(D1):D24–D33. doi:10.1093/nar/gkz913.
21. Xu C, Chen YP, Du XJ, Liu JQ, Huang CL, Chen L, Zhou GQ, Li WF, Mao YP, Hsu C, et al.. Comparative safety of immune checkpoint inhibitors in cancer: systematic review and network meta-analysis. BMJ 2018;363:k4226. doi:10.1136/bmj.k4226.
22. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an experimental factor ontology. Bioinformatics. 2010;26(8):1112–1118. doi:10.1093/bioinformatics/btq099.
23. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al.. DrugBank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–D1082. doi:10.1093/nar/gkx1037.
24. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40(1):30–43. doi:10.1016/j.jbi.2006.02.013.
25. Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–D515. doi:10.1093/nar/gky1049.
26. Giglio M, Tauber R, Nadendla S, Munro J, Olley D, Ball S, Mitraka E, Schriml LM, Gaudet P, Hobbs ET, et al.. ECO, the evidence & conclusion ontology: community standard for evidence information. Nucleic Acids Res. 2019;47(D1):D1186–D1194. doi:10.1093/nar/gky1036.
27. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N; GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–585. doi:10.1038/ng.2653.
28. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–1120. doi:10.1038/ng.2764.
29. Yu Y, Wang Y, Xia Z, Zhang X, Jin K, Yang J, Ren L, Zhou Z, Yu D, Qing T, et al.. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Res. 2019;47(D1):D1090–D1101. doi:10.1093/nar/gky1042.
30. Zhang W, Zhang H, Yang H, Li M, Xie Z, Li W. Computational resources associating diseases with genotypes, phenotypes and exposures. Brief. Bioinform. 2019;20(6):2098–2115. doi:10.1093/bib/bby071.
31. Zhang W, Yao G, Wang J, Yang M, Wang J, Zhang H, Li W. ncRPheno: a comprehensive database platform for identification and validation of disease related noncoding RNAs. RNA Biol. 2020;17(7):943–955. doi:10.1080/15476286.2020.1737441.
32. Hagen NT, DeSalle R. Harmonic allocation of authorship credit: source-level correction of bibliometric bias assures accurate publication and citation analysis. PLoS One. 2008;3(12):e4021. doi:10.1371/journal.pone.0004021.
33. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, Stein LD, Ferretti V. The international cancer genome consortium data portal. Nat Biotechnol. 2019;37(4):367–369. doi:10.1038/s41587-019-0055-9.
34. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al.. NCBI GEO: archive for functional genomics data

sets--update. Nucleic Acids Res. 2013;41(Database issue):D991–D995. doi:10.1093/nar/gks1193.
35. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, et al.. IMGT(R), the international immunogenetics information system(R) 25 years on. Nucleic Acids Res. 2015;43(D1):D413–D422. doi:10.1093/nar/gku1056.
36. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SIPD. IMGT/HLA database. Nucleic Acids Res. 2020;48(D1):D948–D955. doi:10.1093/nar/gkz950.
37. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019;47(D1):D339–D343. doi:10.1093/nar/gky1006.