SOFTWARE TOOL ARTICLE

Cancer Publication Portal: an online tool for summarizing and searching human cancer-genomic publications [version 1; peer review: 2 approved with reservations]

Garrett M. Dancik¹, Kevin Williams², Myron Zhang², Nataliia Romanenko²

¹Department of Computer Science, Eastern Connecticut State University, Willimantic, CT, 06626, USA

²Program in Computer Science, Eastern Connecticut State University, Willimantic, CT, 06626, USA

V1 First published: 10 Dec 2019, 8:2073
<https://doi.org/10.12688/f1000research.21463.1>

Latest published: 10 Dec 2019, 8:2073
<https://doi.org/10.12688/f1000research.21463.1>

Abstract

A search of PubMed lists >582,000 citations with the keywords “cancer” and “gene”. The large volume of cancer genomic publications necessitates the development of text-mining tools to help cancer researchers navigate and summarize articles efficiently. We developed a Cancer Publication Portal (CPP) to help researchers efficiently search and summarize cancer genomic publications, based on one or more genes of interest. CPP integrates data from several sources, including PubTator, the Medical Subject Headings (MeSH) database; the HUGO Gene Nomenclature Committee human gene name database; PubMed, a database of biomedical literature citations; and the National Cancer Institute (NCI) Thesaurus. Following each query, results are summarized and include the publication frequency for each cancer type, as well as publication frequencies for cancer terms, pharmacological agents, genomic mutations, and additional genes stratified by cancer type. Cancer terms were identified by comparing titles and abstracts from cancer-related (N=851,868) and non-cancer related articles (N=2,607,020). CPP allows a user to quickly obtain publication statistics, such as the frequency of articles mentioning *EGFR* across cancer types, and to explore associations, such as the association between pharmacological agent and cancer type. Result summaries are interactive, so additional filters can be easily added as the literature is explored. After a search is completed, a PubTator collection can be quickly created, in order to view article titles and abstracts in PubTator. CPP currently includes information for ~1.1 million cancer-related publications associated with >23,000 human genes.

Database URL: <https://gdancik.github.io/bioinformatics/CPP/>.

Keywords

Text-mining, Cancer

Open Peer Review

Approval Status ? ?

	1	2
version 1	?	?
10 Dec 2019	view	view

1. **Elspeth A. Bruford** , European Bioinformatics Institute (EMBL-EBI), Hinxton, UK
University of Cambridge, Cambridge, UK
2. **Qingyao Huang**, University of Zurich, Zürich, Switzerland
Damian Szklarczyk, University of Zurich, Zürich, Switzerland
SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Garrett M. Dancik (dancikg@easternct.edu)

Author roles: **Dancik GM:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Williams K:** Formal Analysis, Software, Writing – Review & Editing; **Zhang M:** Software, Writing – Review & Editing; **Romanenko N:** Software, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported, in part, by a grant from the American Association of University Professors and Connecticut State University Board of Regents (DANR18).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Dancik GM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Dancik GM, Williams K, Zhang M and Romanenko N. **Cancer Publication Portal: an online tool for summarizing and searching human cancer-genomic publications [version 1; peer review: 2 approved with reservations]** F1000Research 2019, 8:2073 <https://doi.org/10.12688/f1000research.21463.1>

First published: 10 Dec 2019, 8:2073 <https://doi.org/10.12688/f1000research.21463.1>

Introduction

Cancer is a genetic disease¹, with relevant genes often identified through functional screening^{2–4}, gene expression profiling^{5–7}, or genomic sequencing experiments^{8–10}. While researchers may need to quickly understand the published literature regarding a particular gene, the large volume of publications can make this challenging. Indeed, a search of PubMed finds >582,000 citations with the keywords “cancer” and “gene”, with >175,000 articles published within the previous 5 years. The large volume of cancer genomic publications necessitates the development of tools to help cancer researchers navigate and summarize articles efficiently.

The use of controlled vocabularies and text-mining tools has facilitated the annotation and searching of biomedical literature. In particular, the National Library of Medicine’s Medical Subject Headings provides a controlled vocabulary of MeSH terms for indexing MEDLINE/PubMed articles. PubTator is a web-based platform, designed to assist biocuration, that uses robust text-mining tools to annotate PubMed articles with respect to genes, chemicals, species, and mutations¹¹. However, while PubTator allows users to search PubMed based on these biological concepts, summaries of the results are not provided. Other tools such as Anne O’Tate¹² and PubReminer¹³ summarize PubMed searches, but are not cancer-specific and have limitations regarding the number of results that can be returned. Anne O’Tate, for example, allows users to search PubMed and summarizes the results based on important words, phrases, authors, and other fields¹². PubReminer¹³ also allows PubMed queries and summarizes articles based on common words, MeSH terms, and other fields. These summaries are useful but are not cancer specific, and cancer type mentions can be difficult to find and may not appear in the search results.

Here we develop a Cancer Publication Portal (CPP), a web application to help users search and summarize the cancer genomic literature^{14,15}. CPP allows a user to enter a human gene or gene set of interest, and then summarizes the relevant cancer-related publications mentioning this gene through tabular and graphical summaries showing the frequency of articles by cancer type, pharmacological agent, genomic mutation, and additional human genes mentioned in article titles or abstracts. Additionally, CPP catalogs and summarizes articles based on mentions of >30 cancer-related terms. The tool is designed to provide users with publication statistics, such as the number of articles mentioning EGFR and erlotinib across cancer types, as well as to facilitate exploration and retrieval of relevant articles. For example, a user can quickly find all articles based on a set of genes and a collection of cancer types. Interactive summaries allow users to narrow in on a topic of interest by applying additional filters. Results are summarized across cancer types, and the process can be repeated. At any point, the user can access article abstracts from PubTator, as well as download statistical summaries. In this fashion, researchers and students can explore the literature and find articles in a gene- and cancer-focused way.

Methods

Design and implementation

CPP^{14,15} integrates data from a variety of sources including PubTator¹¹, the Medical Subject Headings (MeSH) database; the HUGO Gene Nomenclature Committee (HGNC) human gene name database¹⁶, the PubMed database of biomedical literature citations, and the National Cancer Institute (NCI) Thesaurus¹⁷. An overview of data collection is provided in Figure 1A. Generally, article association data for genes, chemicals, diseases, and mutations was collected from PubTator, then updated and filtered as described below and in Figure 1A. Cancer term associations were identified by comparing PubMed abstracts, as described below and in Figure 1A. Summary statistics for CPP, which contains article-entity association information for 1,143,191 articles and 19,551 genes, is provided in Table 1. The mean number of articles per gene is 138.6 (median = 9.0), but the number of publications per gene is uneven, with the 192 most frequently mentioned genes accounting for >50% of all publications. The three most frequently mentioned genes are *TNF* (N = 64,103), *TP53* (62,602), and *EGFR* (35,178) (Extended data, Supplementary Table S1)¹⁸. The three most common cancer types are Breast Neoplasms (N = 119,891), Leukemia (N = 103,222), and Lymphoma (N = 60,320) (Extended data, Supplementary Table S2)¹⁸.

Collection and processing of PubTator data. Data defining article-gene, article-disease, article-chemical, and article-mutation relationships were downloaded from PubTator via FTP. PubTator data defines associations between articles and mentions of genes, chemicals, diseases, and mutations in article titles or abstracts. MeSH terms for descriptor and supplemental record sets were downloaded as XML files and parsed to extract MeSH IDs and their corresponding terms. A list of pharmacologically active compounds was also downloaded from MeSH via its FTP service. PubTator data was filtered to include only those articles mentioning human genes and only those articles that are cancer-related, i.e., that mention MeSH terms falling under the heading “Neoplasm” (Tree Number C04). We take advantage of MeSH tree structure and remove redundant MeSH IDs if a child MeSH ID is mentioned in the same article. For example, an article mentioning both “Breast Neoplasms” (C04.588.180) and “Triple Negative Breast Neoplasms” (C04.588.180.788) would have the former removed. We also recode “Neoplasms” (C04) as “Neoplasms (unspecified)” (C04.000) if this is the only cancer MeSH term associated with the article. Obsolete MeSH IDs were updated by testing the terms associated with that entry against current MeSH headings and Supplementary Concept record terms. Mutation data was reformatted according to HGVS sequence variant nomenclature recommendations¹⁹. For each mutation, we identify the gene or genes most commonly associated with it, and store this information in CPP.

Identification and annotation of cancer terms. In addition to the article associations provided by PubTator, we report associations between articles and “cancer terms”. We identified

Figure 1. Overview of Cancer Publication Portal (CPP) construction. (A) CPP integrates data from PubTator, PubMed, HGNC, MeSH, and the NCI Thesaurus to summarize articles based on their references to cancer type, mutations, genes, and cancer-related terms. (B) Selected cancer-related terms identified from the titles/abstracts of ~3.5 million publications. The log₁₀ ratio of the cancer publication frequency to non-cancer publication frequency is shown.

Table 1. Cancer Publication Portal statistics.

Variable	Frequency	
	Entity	Article-entity associations
Genes	19,551	2,710,512
Cancer types	660	1,626,382
Cancer terms	37	4,571,767
Pharmacological agents	5,565	726,501
Mutations	32,854	91,831
Articles	1,143,191	-

cancer-terms by comparing title/abstract word ‘stems’ between cancer-related (N=851,868) and non-cancer related articles (N=2,607,020) that mentioned at least one human gene. Abstracts were downloaded from PubMed’s FTP service. These titles/abstracts contained a total of 5,564 unique word stems, with 2,633 word stems more common in cancer articles ($P < 0.01$, Fisher’s exact test). In order to focus on word stems that would be most informative in a cancer-specific context, we filtered these results by considering only word stems that occurred in $> 1\%$ of cancer-related articles. Word stems related to disease/tissue (e.g., ‘renal’), gene name (e.g., ‘kinase’), and

miscellaneous words (e.g., ‘report’) were also filtered out. Word stems for similar words were combined, based on common word usage and the NCI Thesaurus¹⁷. In addition, several terms deemed important by the authors (e.g., “immunotherapy”) were added, even if occurring $< 1\%$ of the time in cancer-related articles. The full list of 37 cancer terms can be seen in the *Extended data*, Supplementary Table S3¹⁸. Selected terms are shown in **Figure 1B**.

For each cancer term, we find a non-redundant set of word stems corresponding to the term itself and its synonyms according to the NCI thesaurus (*Extended data*, Supplementary Table S3)¹⁸. Such an approach allows us to identify a concept (e.g., ‘mutation’) even when a synonym (e.g., ‘genetic alteration’) is used. For each cancer-related article, we search its title/abstract for mentions of cancer terms and add these relationships to the CPP database.

Technical details. Python and R v3.5.2 were used for data processing. PubMed files were parsed using the Python “PubMed Parser”²⁰ and word stems found using the Snowball stemmer from Python’s NLTK module, after removal of stop words and any word with no more than 3 characters. Additional XML files were parsed using the Python module *lxml*. After processing, data was loaded into a MySQL database. The web interface was developed using R/Shiny v1.2.0.

Operation

CPP runs in a standard web browser and is available for public use at the following address: <https://gdancik.github.io/bioinformatics/CPP/>.

Use cases

Summarizing cancer types in publications that mention EGFR

CPP^{14,15} takes a gene-centric approach for finding and summarizing cancer-related articles based on mentions of cancer types, cancer terms, drugs, mutations, and additional genes. In order to demonstrate the utility of CPP, we use CPP to summarize and explore cancer-related articles mentioning the gene epidermal growth factor receptor (*EGFR*), a gene mutated in >30% of patients with non-small cell lung cancer²¹ and a gene that can be targeted by tyrosine kinase inhibitors such as gefitinib and erlotinib²². The user starts by selecting the desired gene from the drop-down menu. After selecting *EGFR*, CPP tells us that there are 35,178 articles found, covering 422 cancer types (Figure 2A). We note that “cancer types” here is defined according to MeSH subject headings, which categorize cancers by both site and histological type. The top three cancer types are “Neoplasms, Glandular and Epithelial”, “Thoracic Neoplasms”, and “Lung Neoplasms”. We next select the cancer types to search, by either clicking on the table or selecting cancer types from the drop-down menu. Cancer types can also be uploaded from a file. After clicking on the button to retrieve the summaries, we get a tabular and graphical summary showing the number of articles mentioning both the selected gene and each selected cancer type (Figure 2B). If no cancer type is selected, then all cancer types will be summarized. Here it is easy to compare the number of articles mentioning *EGFR* across cancer types, and a user quickly sees that lung cancer is the most common.

Summarizing drug, mutation, and gene mentions in EGFR cancer-related publications

Figure 3 shows additional summaries provided by CPP. Summaries include frequency tables showing the number of articles associated with the selected gene and Cancer Terms, Drugs, Mutations, and Additional Genes; and stacked bar graphs for visualizing the distribution of each entity across cancer types. If the user searched for multiple genes, a summary of the selected genes is also provided. The frequency tables allow a user to quickly identify entities (such as drugs) that are commonly mentioned in the literature, while the stacked bar graphs allow a user to qualitatively evaluate when entities are more (or less) associated with specific cancer types than others. In this example, frequency tables show that gefitinib and erlotinib are the two drugs most commonly associated with *EGFR* (Figure 3A), the *EGFR* mutations p.T790M and p.L858R are most common (Figure 3B), and *ERBB2*, *KRAS*, and *TP53* are the genes that most frequently co-occur with *EGFR* (Figure 3C). However, in the latter case the stacked bar graph shows that these co-occurring genes are associated with specific cancer types. Specifically, while *KRAS* is the most common gene that co-occurs with *EGFR* in lung cancer, the most common genes that co-occur with breast cancer and glioma are *ERBB2* and *TP53*, respectively (Figure 3C). Such associations may reflect

genomic differences between cancer types, or may reflect a bias in the literature. The stacked bar graphs are interactive. A user can single click on an entity in the legend to hide the entity from the graph, and double-click on an entity to hide all other entities. This toggling can be canceled by clicking or double clicking the entity a second time. For example, by double clicking on the drug irinotecan, we can see that this drug is associated with colorectal cancers more than other cancer types (Figure 3A, inset).

Finding articles assessing the predictive value of EGFR mutations for gefitinib or erlotinib treatment

In addition to summarizing the cancer genomic literature, CPP is designed to help users explore the literature and quickly find articles of interest. After selecting one or more genes and cancer types, a user can add filters by clicking on one or more rows of any frequency table to add an entity to the filter. For each entity type, the user can retrieve articles that mention either all of the selected terms or any of the selected terms. For example, a user interested in publications assessing the predictive value of *EGFR* mutations as biomarkers for gefitinib or erlotinib treatment could use CPP to first find articles that mention *EGFR* and any cancer type. Then the user can specify additional filters to limit the results to articles that mention both mutation and survival, and either of the drugs gefitinib or erlotinib (Figure 4A). Note that cancer term filters are based on word ‘stems’ and synonyms from the NCI Thesaurus, and therefore will recognize variations of the search term. For example, the cancer term “mutation” includes any words with a stem of ‘mutat’, which includes the words “mutation”, “mutations”, and “mutated”; and word stems corresponding to ‘genetic alteration’ and ‘genetic change’ (Extended data, Supplementary Table S3)¹⁸. This search results in 1,169 articles being found. In summarizing the articles, we see that lung cancers are the dominant cancer type (Figure 4B), and that the number of articles is similar for each drug across cancer types (Figure 4C). We note that the stacked bar graphs are now limited to the entities we have selected (i.e., gefitinib and erlotinib).

Users can easily explore the literature by adding and removing filters. At any point, a user can view abstracts for the current set of articles. A user views abstracts by selecting the ‘Articles’ tab, clicking the ‘Copy PMIDs’ button, and then creating a new collection in PubTator, which is displayed on the Articles page using an iframe (Figure 4D). This allows a user to seamlessly view relevant abstracts after applying the desired filters. Additionally, a user can download results to a CSV file, for the current list of PMIDs, as well as frequency summaries for Cancer Types, Cancer Terms, Drugs, Mutations, and Additional Genes from the ‘Download’ tab.

Discussion and conclusions

CPP^{14,15} is designed to help users efficiently explore and summarize the cancer genomic literature, and should be useful for cancer researchers who are looking for relevant articles for a gene of interest, for meta-researchers who study the publication landscape, and for students learning about the relationship between one or more genes and cancer types. Because CPP

Figure 2. Cancer Publication Portal screenshots for gene and cancer type selection. (A) Cancer selection screen displayed after a user enters one or more genes. **(B)** Cancer types summary screen, showing frequency table and bar graph of selected cancer types.

Figure 3. Cancer Publication Portal screenshots summarizing article associations. Summaries and stacked bar graphs are provided for Cancer terms (not shown), (A) Drugs, (B) Mutations, and (C) mentions of additional Genes. Inset in (A) shows only irinotecan, obtained by double clicking on that drug in the legend.

Figure 4. Cancer Publication Portal screenshots demonstrating filtering and abstract viewing. (A) Filters can be specified for all selected terms for an entity or any selected term. Current filter shows articles mentioning mutation and survival and either gefitinib or erlotinib. (B) Cancer summary of results for EGFR, all cancer types, and the filters in (A). (C) Stacked bar graph showing mentions of gefitinib and erlotinib across cancer types. (D) Screenshot of 'Articles' tab where user can create a PubTator collection to view the current set of articles.

quickly summarizes articles across cancer types, CPP can be used to assess whether a gene might be novel for a particular cancer type (or any cancer type), based on the frequency of gene mentions in titles/abstracts of cancer-related publications. CPP also provides summaries of cancer terms, drugs, mutations, and co-occurring genes that can connect a researcher to key biological concepts and the underlying articles that might inform their research. The use of cancer terms, unique to CPP, summarizes articles in a cancer-specific way and allows for quick retrieval of articles based on a cancer term, such as tumor suppressor, chemotherapy, or metastasis. Furthermore, because CPP identifies associations between articles and cancer terms by using the NCI Thesaurus and ‘stem’ words to cover synonyms and variant word forms (such as ‘mutation’ and ‘mutated’), CPP will retrieve a more valid set of articles than a simple PubMed search for a term in other databases.

Despite its utility, CPP has several limitations that are common in all text-mining based tools. With the exception of cancer terms, article associations in CPP are derived from PubTator associations. While some associations may be missed, PubTator uses cutting edge text-mining tools and the F_1 scores for gene, disease, and mutation identification are all > 80%¹¹.

Importantly, text-mining associations are determined only by textual relationships and may not reflect underlying biology. For example, genes that are mentioned together in the same abstract may or may not interact. Similarly, publication frequency may reflect publication bias and not biological importance. Other tools are available for looking at specific biological relationships, such as STRING for protein interactions²³, and cBioPortal for Cancer Genomics for exploring genomic datasets²⁴. CPP is designed to complement these tools by providing an overview of the cancer genomic literature and by helping researchers quickly find relevant publications. Finally, all CPP associations, which are derived from PubTator, are based on entity mentions in titles and abstracts only. However, the recently released PubTator Central includes associations from the PubMed Central (PMC) Text Mining Subset²⁵. PMC contains the full text of ~3 million articles, though we expect <4% to be cancer-related. PubTator Central associations will be integrated into CPP in a future release.

In conclusion, CPP is an easy-to-use web application that allows researchers to efficiently summarize and search the cancer literature for articles based on one or more genes of interest. CPP will be updated approximately once a month following PubTator data releases.

Data availability

Source data

Associations between genes, diseases, chemicals, mutations, and articles were downloaded from the PubTator FTP page (<ftp://ftp.ncbi.nlm.nih.gov/pub/lu/PubTator/>)

MeSH descriptor files were downloaded from <https://www.nlm.nih.gov/databases/download/mesh.html>

Extended data

Dataverse: Cancer Publication Portal: an online tool for summarizing and searching human cancer-genomic publications (supporting data). <https://doi.org/10.7910/DVN/BYKFIL>¹⁸

This project contains the following extended data:

- **Supplementary Table S1. Number of articles per gene in Cancer Publication Portal.**
- **Supplementary Table S2. Number of articles per cancer type in Cancer Publication Portal.** Cancer types are defined by cancer-related MeSH TreeIDs (C04*).
- **Supplementary Table S3. Cancer terms and synonyms included in Cancer Publication Portal.** For each term, there was a statistically significant difference in the proportion of mentions between cancer-related and non-cancer related articles ($P < 0.001$ by Fisher’s exact test). *, term is included despite appearing in < 1% of cancer-related articles, and/or not being cancer-specific (i.e., $\log_{10}$ ratio < 1).

Extended data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

Software availability

CPP is available as a web resource at: <https://gdancik.github.io/bioinformatics/Cpp/>.

Source code for the web interface is available from: <https://github.com/gdancik/Cpp>.

Source code for data processing and database creation is available from: https://github.com/gdancik/Cpp_setup.

The database is available from the following docker image: <https://hub.docker.com/r/gdancik/dcast>.

Archived source code for web interface at time of publication: <https://doi.org/10.5281/zenodo.3550110>¹⁴.

Archived source for data processing and database creation at time of publication: <https://doi.org/10.5281/zenodo.3550112>¹⁵.

License: [GNU General Public License-2](#).

Acknowledgements

The authors acknowledge Stefanos Stravoravdis for coding contributions, and Andrew Johnson for coding contributions and technical assistance. The authors also acknowledge Jason Duex and Sunny Guin for testing and providing feedback for an earlier version of the tool.

References

1. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome**. *Nature*. 2009; 458(7239): 719–24.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Guin S, Pollard C, Ru Y, *et al.*: **Role in tumor growth of a glycogen debranching enzyme lost in glycogen storage disease**. *J Natl Cancer Inst*. 2014; 106(5): pii: dju062.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Han R, Li L, Ugalde AP, *et al.*: **Functional CRISPR screen identifies AP1-associated enhancer regulating FOXF1 to modulate oncogene-induced senescence**. *Genome Biol*. 2018; 19(1): 118.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Falkenberg KJ, Newbold A, Gould CM, *et al.*: **A genome scale RNAi screen identifies *GLI1* as a novel gene regulating vorinostat sensitivity**. *Cell Death Differ*. 2016; 23(7): 1209–18.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Youns M, Efferth T, Reichling J, *et al.*: **Gene expression profiling identifies novel key players involved in the cytotoxic effect of Artesunate on pancreatic cancer cells**. *Biochem Pharmacol*. 2009; 78(3): 273–83.
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Lee RS, Zhang L, Berger A, *et al.*: **Characterization of the ERG-regulated Kinome in Prostate Cancer Identifies TNK1 as a Potential Therapeutic Target**. *Neoplasia*. 2019; 21(4): 389–400.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Reyes I, Reyes N, Suriano R, *et al.*: **Gene expression profiling identifies potential molecular markers of papillary thyroid carcinoma**. *Cancer Biomark*. 2019; 24(1): 71–83.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Collins CC, Volik SV, Lapuk AV, *et al.*: **Next generation sequencing of prostate cancer from a patient identifies a deficiency of methylthioadenosine phosphorylase, an exploitable tumor target**. *Mol Cancer Ther*. 2012; 11(3): 775–83.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Labgaa I, Villacorta-Martin C, D'Avola D, *et al.*: **A pilot study of ultra-deep targeted sequencing of plasma DNA identifies driver mutations in hepatocellular carcinoma**. *Oncogene*. 2018; 37(27): 3740–52.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Lee WK, Lee SG, Yim SH, *et al.*: **Whole Exome Sequencing Identifies a Novel Hedgehog-Interacting Protein G516R Mutation in Locally Advanced Papillary Thyroid Cancer**. *Int J Mol Sci*. 2018; 19(10): pii: E2867.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Wei CH, Kao HY, Lu Z: **PubTator: a Web-based text mining tool for assisting Biocuration**. *Nucleic Acids Res*. 2013; 41(Web Server issue): W518–22.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Smalheiser NR, Zhou W, Torvik VI: **Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results**. *J Biomed Discov Collab*. 2008; 3: 2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. **PubReMiner: a tool for PubMed query building and literature mining** [Internet]. [cited 2019 Jun 17].
[Reference Source](#)
14. Dancik G, Johnson A, Romanenko N: **gdancik/CPP: CPP (F1000 release) (Version 1.0)**. *Zenodo*. 2019.
<http://www.doi.org/10.5281/zenodo.3550110>
15. Williams K, Zhang M, Dancik G: **gdancik/CPP_setup: CPP_setup (F1000Research) (Version 1.0)**. *Zenodo*. 2019.
<http://www.doi.org/10.5281/zenodo.3550112>
16. Yates B, Braschi B, Gray KA, *et al.*: **Genenames.org: the HGNC and VGNC resources in 2017**. *Nucleic Acids Res*. 2017; 45(D1): D619–25.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. **NCI Thesaurus** [Internet]. [cited 2019 Jun 20].
[Reference Source](#)
18. Dancik G, Williams K, Zhang M, *et al.*: **Cancer Publication Portal: an online tool for summarizing and searching human cancer-genomic publications (supporting data)**. *Harvard Dataverse*, V1, UNF:6.5POzQ6fu7p4qBw5J6vIFpQ== [fileUNF]. 2019.
<http://www.doi.org/10.7910/DVN/BYKFL1>
19. den Dunnen JT, Dalgleish R, Maglott DR, *et al.*: **HGVS Recommendations for the Description of Sequence Variants: 2016 Update**. *Hum Mutat*. 2016; 37(6): 564–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Achakulvisut T, Acuna DE: **PubMed Parser** [Internet]. *PubMed Parser*. [cited 2015 Jul 2]. 2015.
[Reference Source](#)
21. Zhang YL, Yuan JQ, Wang KF, *et al.*: **The prevalence of *EGFR* mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis**. *Oncotarget*. 2016; 7(48): 78985–93.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Rocha-Lima CM, Soares HP, Raez LE, *et al.*: **EGFR targeting of solid tumors**. *Cancer Control*. 2007; 14(3): 295–304.
[PubMed Abstract](#) | [Publisher Full Text](#)
23. Szklarczyk D, Franceschini A, Wyder S, *et al.*: **STRING v10: protein-protein interaction networks, integrated over the tree of life**. *Nucleic Acids Res*. 2015; 43(Database issue): D447–452.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Gao J, Aksoy BA, Dogrusoz U, *et al.*: **Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal**. *Sci Signal*. 2013; 6(269): pi1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Wei CH, Allot A, Leaman R, *et al.*: **PubTator central: automated concept annotation for biomedical full text articles**. *Nucleic Acids Res*. 2019; 47(W1): W587–93.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 21 February 2020

<https://doi.org/10.5256/f1000research.23643.r59613>

© 2020 Szklarczyk D et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Qingyao Huang

Institute of Molecular Life Sciences, University of Zurich, Zürich, Switzerland

Damian Szklarczyk

¹ Institute of Molecular Life Sciences, University of Zurich, Zürich, Switzerland

² SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

The publication describes the web tool for search, retrieval and computing statistics of cancer publications related to the input gene(s). The tool fully builds upon open source data from PubTator which links medical and biological concepts to PubMed articles. The only derived data is the "Cancer Terms" associations data set. These are words (tags) that are significantly associated with the cancer related publications - that is - publications mentioning any cancer MeSH term (and at least one gene) significantly more relative to the all other publication (with at least one gene mentioned). Methodology is adequately described and the code is publicly available.

Interaction with the website:

- The user queries the database with a gene or a set of genes, and is presented with a list of cancer types associated with their list sorted by a number of associations. After the user selected the cancer type(s) of interest, it is presented with a bar graph with frequencies (counts) of these cancer types in the literature. Furthermore the user can filter the results by Cancer Terms, Drugs, Mutations and Additional Genes. Each of these filters allows the user to create a simple bar graph with the 10 most frequent term on one axis and number of articles on the other. The site is fast and responsive even for large gene list. All results can be downloaded and the code is publicly available.

Main issues/limitation:

- The main issue with the tool is lack of any statistical testing. The web tool only lists frequencies (counts) of these associations, which is not particularly informative. The tool should provide information if the user's initial gene or gene set is significantly more associated with particular cancer types/terms/drugs and how specific this association is, this results in an unspecific terms being always listed first. More crucially however, for example, a drug could have 10 associations with a user input gene out of a total of 10 association in the full corpus, or it could be 10 out of 10.000 in total. The interpretation of

the results would substantially differ in both of these cases, however with the current state of the web tool, there is way to tell these two cases apart.

- Another glaring limitation of the tool is that the only entry point is a gene or a gene set. There seem to be no good reason why the initial query would not be a cancer type, cancer term or the a drug. It's would be a valid and potentially interesting question to ask the tool: Give me all the cancer types and genes associated with, for example, Roscovitine. As of now there is no way to generate such a data set as the output is limited to the associations with the entry gene-list.

Other Issues/limitations:

- The input gene selection box stops listing genes at the letter "C". Also the list is not properly sorted, there are rare other random genes beginning on T or S in between the sorted genes.
- Web browsers back button doesn't function properly (I guess it's partially a limitation of the framework used, but the developers should avoid frameworks and tools that break basic browser UI functions - at very least the tool should warn the user that they work will be lost when the back button is pressed).
- The user can't share the state of the website (their results) with other users.
- When I click on "New Gene Search" I can't select the same gene.
- When I double click on the cancer type in the "Select Cancer Types" window couple of times, I can force the app infinitely starts refreshing.
- "Cancer publication Portal" should be a hyperlink which sends the user to the start page.
- iFrames, such as the ones used for PubTator should be avoided, except looking out of place and being confusing to the user, it's an unstable solution as it may break functionality of framed websites such as pop-ups, full screen features (e.g. videos), and back button. In addition to that not all we browsers support iFrames, it's regarded as unsafe, it breaks the webpage for the impaired users, and one cannot copy/paste the URL of the iframe which is a fundamental usability issue. Possible solution is to link directly to the search results. PubTator search result have simple GET requests scheme which could easily be implemented (the authors should be aware that there is an URL size limit and how many PMID PubTator can actually process, but these limitations also applies to the current solution).
- The authors should consider text-mining NCIt Neoplasm Core instead relying in PubTator MeSH terms. The web tool focus is on oncology and cancer researchers. The Terminology provided by MeSH system is known for lacking granularity. NCIt has extensive hierarchy for cancer-related terms with high coverage. Utilizing inferior ontology makes the tool less useful to the target audience.
- PubTator is already an outdated tool and is now developed further as Pubtator 2.0, which includes, among other improvements, more sophisticated text-mining and corpus of full text papers. The authors should move their pipeline to PubTator 2.0 system.

Is the rationale for developing the new software tool clearly explained?

No

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Protein-protein interaction prediction, protein orthology, text-mining, statistical analysis of experimental data, web/tool development.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 15 Jul 2020

Garrett M. Dancik, Eastern Connecticut State University, Willimantic, USA

I want to thank the reviewers for many constructive comments and for thoroughly testing the *CPP* tool. While we will not be able to formally revise and resubmit the manuscript at this time, we have made some minor changes to the tool and address some of the reviewer's comments below:

Comment: The main issue with the tool is lack of any statistical testing. The web tool only lists frequencies (counts) of these associations, which is not particularly informative. The tool should provide information if the user's initial gene or gene set is significantly more associated with particular cancer types/terms/drugs and how specific this association is, this results in an unspecific terms being always listed first. More crucially however, for example, a drug could have 10 associations with a user input gene out of a total of 10 association in the full corpus, or it could be 10 out of 10.000 in total. The interpretation of the results would substantially differ in both of these cases, however with the current state of the web tool, there is way to tell these two cases apart.

Response: We appreciate this comment and have updated our tool to calculate and rank results by statistical significance. Specifically, an enrichment score is calculated for each term (cancer type, drug, etc), which measures how much more likely a term appears in the selected articles compared to all articles in the database. For example, a *score* of 4 means that the term is 4x as likely to appear in the title/abstract of selected articles than all cancer-related articles in the database. P-values, FDRs, and additional statistical information are provided for each set of results under the "Full Table" tab.

Comment: Another glaring limitation of the tool is that the only entry point is a gene or a gene set. There seem to be no good reason why the initial query would not be a cancer type, cancer term or the a drug. It's would be a valid and potentially interesting question to ask the tool: Give me all the cancer types and genes associated with, for example, Roscovitine. As of now there is no way to generate such a data set as the output is limited to the associations with the entry gene-list.

Response: We acknowledge that this is a limitation of the tool, which was designed to be gene-centric in nature. The tool is appropriate for users wanting to summarize cancer-related articles containing one or more genes.

Comment: The input gene selection box stops listing genes at the letter "C". Also the list is not properly sorted, there are rare other random genes beginning on T or S in between the sorted genes.

Response: In our previous version of *CPP*, we had mistakenly sorted the genes by GeneID, rather than alphabetically by gene symbol. We have corrected this and now sort genes alphabetically. While the input box does not show all genes, the user can start typing into the text box to retrieve matching genes. This feature is now stated explicitly in the drop down label.

Comment: Web browsers back button doesn't function properly (I guess it's partially a limitation of the framework used, but the developers should avoid frameworks and tools that break basic browser UI functions - at very least the tool should warn the user that they work will be lost when the back button is pressed). The user can't share the state of the website (their results) with other users.

Response: We acknowledge that these are current limitations of the tool, and are important features that may be incorporated in the future. We have added a note to the user on the welcome page that the Back button is not functional on the page.

Comment: When I click on "New Gene Search" I can't select the same gene.

Response: This is intentional in order to reduce the computational burden on our server; if a user wants to "reset" a search, the user can clear the filters or click the "Cancer Publication Portal" link to refresh the page (see next item).

Comment: "Cancer publication Portal" should be a hyperlink which sends the user to the start page.

Response: Done

Comment: iFrames, such as the ones used for PubTator should be avoided, except looking out of place and being confusing to the user, it's an unstable solution as it may break functionality of framed websites such as pop-ups, full screen features (e.g. videos), and back button. In addition to that not all we browsers support iFrames, it's regarded as unsafe, it breaks the webpage for the impaired users, and one cannot copy/paste the URL of the iframe which is a fundamental usability issue. Possible solution is to link directly to the search results. PubTator search result have simple GET requests scheme which could easily be implemented (the authors should be aware that there is an URL size limit and how many PMID PubTator can actually process, but these limitations also applies to the current solution).

Response: We agree that iframes have limitations, but prefer them in this case since it makes viewing of search results easier. In addition, we provide a link to PubTator Central so that users can access the page directly if they prefer.

Comment: PubTator is already an outdated tool and is now developed further as Pubtator 2.0, which includes, among other improvements, more sophisticated text-mining and corpus of full text papers. The authors should move their pipeline to PubTator 2.0 system.

Response: Although moving our pipeline to PubTator 2.0 (PubTator Central) was our intention, as mentioned in the Discussion of our manuscript, we are freezing our tool currently with the final release of PubTator (on 2/15/2020). While PubTator 2.0 offers improvements that are useful, the use of full text papers in our experience limits its usability for our purposes. In particular, text-mining (as of 7/3/2020) may identify terms that are mentioned outside of the main paper (e.g., the references or author contributions section). As a result, we have found the number of gene mentions to be inaccurate, and in some cases by orders of magnitude, based on a testing set of genes.

Competing Interests: No competing interests were disclosed.

Reviewer Report 04 February 2020

<https://doi.org/10.5256/f1000research.23643.r58622>

© 2020 Bruford E. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Elsbeth A. Bruford

1

HUGO Gene Nomenclature Committee (HGNC), European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

² Department of Haematology, University of Cambridge, Cambridge, UK

The authors present a new tool that aims to simplify searching for cancer-related publications. There is no doubt that the number of publications in this field is ever increasing and hence a tool like this could prove useful in narrowing down the number of citations that could be of interest.

Is the description of the software tool technically sound?

I said "partly" as the "technical details" within the manuscript itself number only a total of 11 lines of text and don't really give any description as such - what does "Snowball stemmer" do? What PubMed files were parsed? Etc.

However, at the end of the manuscript the authors state all of the software is freely available in GitHub, so while there is very little discussion of the software tool itself, hopefully someone could reconstruct the work using the software listed (I wasn't about to try this).

I used the URL listed for accessing the tool - in fact this URL then directs you to another URL, <http://bioinformatics.eaternct.edu/app/CPP/>. I have no idea why that URL isn't actually listed in the manuscript, and it was disappointing to see that they haven't bothered to make this https instead of http.

A few comments on the user interface: the home button doesn't seem to take you anywhere - if you've run a search and click home you stay exactly where you are on the results page. Also clicking on "Cancer Publication Portal" doesn't take you anywhere; personally I find it quite irritating that the initial stages of any search, selecting your filters, involve a modal popup window, which then closes, it also makes modifying the terms clunky; the order of the genes listed in the drop-down seems very odd (I think it could be based on the NCBI Gene ID for each gene?) - it would make more sense to make the list alphabetical; when selecting the "terms", it would be helpful to have an "X" next to each selected term to make it clear how they can be removed (clicking delete did remove the term but it also did something odd to the "selected" box, where a dropdown suddenly appeared...(using Firefox 72.0.2 in Windows 10)); variations could be viewed as a better term than "mutations"; in the "mutations".

A key aspect of the tool is that after applying the filters you can then either download the resulting set of PubMed IDs, or you can transfer (by literally copying and pasting) the IDs to view them in "PubTator", which opens in an iframe. I did find it variable whether "PubTator" would retrieve results or I would simply get a blank iframe. Also, some results give 1000s of PMIDs, which is clearly far too many for PubTator to cope with (again, blank iframe resulted). Several times I got results that said:

Welcome! Guest. | [Log in](#)

Results: 1 to 3 of 3

1 [No Title](#)

PMID:{"error":"API - Related citations
ABSTRACT not available

2No Title

PMID:rate - Related citations
ABSTRACT not available

3No Title

PMID:limit - Related citations
ABSTRACT not available

4No Title

PMID:exceeded","api-key":"130.14.18.113","count":"4","limit":"3"} - Related citations
ABSTRACT not available

and that was simply copying and pasting 3 PMIDs (note typo in word "available").

So I am not convinced that the viewing in PubTator aspect, while potentially useful, is actually fully operational. Furthermore, when PubTator opens in the iframe it clearly says "You will be automatically redirected to the new and improved PubTator Central (PubTator 2.0) website after January 2020. " Well, it's definitely February and I wasn't redirected anywhere...

The methodology in general seems to rely heavily upon PubTator; the PubTator 2019 paper claims it is updated daily so it would be good to know how often CPP updates their initial "PubTator" data. Indeed, how often they update any of the data sources and how their update cycles run would be very helpful. I am also confused by Figure 1, as it has PubMed as a separate input from PubTator, but I understood that PubTator was based on PubMed? Surely concept-article associations have to have source articles in the first place?

The abstract states "CPP currently includes information for ~1.1 million cancer-related publications associated with >23,000 human genes" but then Table 1 and the text states 19,551 genes. Also, why 19,551 genes? This is never explained. The set appears to include pseudogenes and long ncRNAs so this could also be worth mentioning.

I am also less than convinced about the usefulness of the "cancer terms" when they can be as general as "patient", "DNA", "diagnosis" etc, but I guess it is up to the user to assess their utility themselves. A lot of emphasis is placed on their inclusion in this tool, but in reality I am dubious about how useful they would actually be.

I think there are far too many figures included, and mostly screenshots - these need to be condensed to show key information, or removed altogether.

In summary, I think this paper describes a tool that is a good concept, but the execution currently needs some more development and there are clearly some bugs. If these bugs were fixed and perhaps some UX testing done to improve the tool and the manuscript discussed this and dealt with the issues I have raised above, both would be greatly improved.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Human genomics, genetics, comparative genomics, bioinformatics, nomenclature, biomedical resources.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 Jul 2020

Garrett M. Dancik, Eastern Connecticut State University, Willimantic, USA

I want to thank the reviewer for many constructive comments and for thoroughly testing the *CPP* tool. While we will not be able to formally revise and resubmit the manuscript at this time, we have made some minor changes to the tool and address some of the reviewer's comments below:

Comment: I used the URL listed for accessing the tool - in fact this URL then directs you to another URL, <http://bioinformatics.eaternct.edu/app/CPP/>. I have no idea why that URL isn't actually listed in the manuscript, and it was disappointing to see that they haven't bothered

to make this https instead of http.

Response: The URL provided in the manuscript, https://gdancik.github.io/bioinformatics/_CPP/, is a stable URL and the homepage for *CPP*. We intentionally chose *not* to link directly to the tool in our manuscript, because computational demands may require moving the tool itself to a different host. The homepage will always link to the current tool. We also agree that https should have been used instead of http, and we have migrated our tool to https.

Comment: A few comments on the user interface: the home button doesn't seem to take you anywhere - if you've run a search and click home you stay exactly where you are on the results page. Also clicking on "Cancer Publication Portal" doesn't take you anywhere; personally I find it quite irritating that the initial stages of any search, selecting your filters, involve a modal popup window, which then closes, it also makes modifying the terms clunky;

Response: We have made a few changes to the interface that we hope improves the user experience. In place of a "Home" tab we now have a "Search" tab, which eliminates the initial modal popup. Following a search, the results are shown on the "Results" page. The user can go back to the "Search" page at anytime to carry out another search. We have also updated the "Cancer Publication" Portal link so that clicking on it reloads the page.

Comment: the order of the genes listed in the drop-down seems very odd (I think it could be based on the NCBI Gene ID for each gene?) - it would make more sense to make the list alphabetical;

Response: We appreciate this comment. The genes are now sorted alphabetically rather than by NCBI Gene ID.

Comment: When selecting the "terms", it would be helpful to have an "X" next to each selected term to make it clear how they can be removed (clicking delete did remove the term but it also did something odd to the "selected" box, where a dropdown suddenly appeared...(using Firefox 72.0.2 in Windows 10));

Response: We have clarified in the instructions that terms may be removed by clicking on the table or by removing them from the dropdown box.

Comment: A key aspect of the tool is that after applying the filters you can then either download the resulting set of PubMed IDs, or you can transfer (by literally copying and pasting) the IDs to view them in "PubTator", which opens in an iframe. I did find it variable whether "PubTator" would retrieve results or I would simply get a blank iframe. Also, some results give 1000s of PMIDs, which is clearly far too many for PubTator to cope with (again, blank iframe resulted).

Response: We agree that viewing selected articles in PubTator is a key aspect of our tool, and our goal is for this process to be as seamless as possible. We have encountered similar API errors in the past, both while using *CPP* as well as when using PubTator directly. These

errors are beyond our control. However, we now explicitly instruct users to go directly to PubTator (now PubTator Central) if encountering errors on our page. We also have tested our tool with PubTator Central and have been able to retrieve citations for >10,000 articles without any issues.

Comment: The methodology in general seems to rely heavily upon PubTator; the PubTator 2019 paper claims it is updated daily so it would be good to know how often CPP updates their initial "PubTator" data. Indeed, how often they update any of the data sources and how their update cycles run would be very helpful.

Response: Our original intention was to update our tool approximately once a month following the PubTator bulk data release schedule. This is mentioned at the end of our manuscript. However, PubTator has moved to PubTator 2.0 (PubTator Central), and while moving our pipeline to PubTator 2.0 (PubTator Central) was our intention, we are freezing our tool currently with the final release of PubTator (on 2/15/2020). While PubTator 2.0 offers improvements that are useful, its use of full text papers in our experience limits its usability for our purposes. In particular, text-mining (as of 7/3/2020) may identify terms that are mentioned outside of the main paper (e.g., the references or author contributions section). As a result, we have found the number of gene mentions to be inaccurate based on a testing set of genes.

Comment: I am also confused by Figure 1, as it has PubMed as a separate input from PubTator, but I understood that PubTator was based on PubMed? Surely concept-article associations have to have source articles in the first place?

Response: This should have been more clear in the figure and in the text. Our **Figure 1** provides information about how data is integrated into *CPP*. The gene, mutation, chemical, and disease associations are downloaded from PubTator. While these associations are based on PubMed, PubMed is not our primary source for this data. However, the cancer term mentions are based on PubMed data which is our primary source.

Comment: The abstract states "CPP currently includes information for ~1.1 million cancer-related publications associated with >23,000 human genes" but then Table 1 and the text states 19,551 genes. Also, why 19,551 genes? This is never explained. The set appears to include pseudogenes and long ncRNAs so this could also be worth mentioning.

Response: The >23,000 number was a mistake on our part. The correct number (at the time of the initial publication), was 19,551 genes, which does include pseudogenes, long ncRNAs, and others. We include all genes (molecules assigned an ID by the HGNC) that are associated with at least one cancer publication based on PubTator.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research