Research Paper

Identifying Scientific Project-generated Data Citation from Full-text Articles: An Investigation of TCGA Data Citation

Jiao Li, Si Zheng, Hongyu Kang, Zhen Hou & Qing Qian†

Institute of Medical Information and Library, Chinese Academy of Medical Sciences, Beijing 100020, China

Citation: Jiao Li, Si Zheng, Hongyu Kang, Zhen Hou & Qing Qian (2016). Identifying Scientific Project-generated Data Citation from Full-text Articles: An Investigation of TCGA Data Citation.

Received: Jan. 20, 2016
Revised: Apr. 28, 2016
Accepted: May 15, 2016

Abstract

Purpose: In the open science era, it is typical to share project-generated scientific data by depositing it in an open and accessible database. Moreover, scientific publications are preserved in a digital library archive. It is challenging to identify the data usage that is mentioned in literature and associate it with its source. Here, we investigated the data usage of a government-funded cancer genomics project, The Cancer Genome Atlas (TCGA), via a full-text literature analysis.

Design/methodology/approach: We focused on identifying articles using the TCGA dataset and constructing linkages between the articles and the specific TCGA dataset. First, we collected 5,372 TCGA-related articles from PubMed Central (PMC). Second, we constructed a benchmark set with 25 full-text articles that truly used the TCGA data in their studies, and we summarized the key features of the benchmark set. Third, the key features were applied to the remaining PMC full-text articles that were collected from PMC.

Findings: The amount of publications that use TCGA data has increased significantly since 2011, although the TCGA project was launched in 2005. Additionally, we found that the critical areas of focus in the studies that use the TCGA data were glioblastoma multiforme, lung cancer, and breast cancer; meanwhile, data from the RNA-sequencing (RNA-seq) platform is the most preferable for use.

Research limitations: The current workflow to identify articles that truly used TCGA data is labor-intensive. An automatic method is expected to improve the performance.

Practical implications: This study will help cancer genomics researchers determine the latest advancements in cancer molecular therapy, and it will promote data sharing and data-intensive scientific discovery.

Originality/value: Few studies have been conducted to investigate data usage by government-funded projects/programs since their launch. In this preliminary study, we extracted articles

JDIS
Journal of Data and Information Science
Vol. 1 No. 2, 2016
pp 32–44
DOI: 10.20309/jdis.201612

† Corresponding author: Qing Qian (E-mail: qian.qing@imicams.ac.cn).

that use TCGA data from PMC, and we created a link between the full-text articles and the source data.

Keywords Scientific data; Full-text literature; Open access; PubMed Central; Data citation

1 Introduction

The scientific community benefits from data sharing. By using previous research data, researchers can advance scientific discovery far beyond their original analysis (Piwowar & Vision, 2013). Scientific data usage facilitates original result confirmation, and it improves new hypothesis generation when combining other types of data.

Biomedical data sharing policies have been established to ensure that data are publicly available. For example, it is mandatory to share large scale genomic data that were generated or analyzed on basis of the U.S. National Institute of Health funding (Green et al., 2015). All of the grantees consciously deposit their data to a public database, and they serve as the data author. To make a collection of enduring scientific data and create a sustainable data ecosystem, all of the key players of data management, such as data author, data curator, data user, and funding agencies, actively fulfill their responsibilities (Bourne, Lorsch, & Green, 2015; National Science Board, 2005). For this data management lifecycle, data authors conform to the data standard and data quality requirement, and they produce scientific data that are further deposited into a public database in a comprehensive manner. Moreover, data users adhere to license or copyright requirements regarding the usage of the data that is generated by data authors, and they must correctly cite the data used in their scientific publications to indicate that their studies feature the use of other research data. Identifying data citations is important for funding agencies to evaluate grantees' scientific contribution and grant outcomes. Additionally, a dataset that is more frequently cited by other researchers is confirmed to be high-quality data and represents domain trends.

It is challenging to identify data citations in full-text literature, although there is a long tradition of partnership between scientific literature and public data in the field of medical sciences (Kafkas, Kim, & McEntyre, 2013). A few studies have been conducted to identify data citation using unique data accession numbers in the database. However, most project-generated data focuses on one specific scientific goal and lacks either a well-defined data identifier or standardized citation regulations.

The Cancer Genome Atlas (TCGA) project was launched in 2005 and funded by the US government, and it aims to catalogue and discover major cancer-causing

genomic alterations to help improve the clinical outcome of cancers (Tomczak, Czerwinska, & Wiznerowicz, 2015). A major goal of the project was to provide publicly available cancer genomic datasets (https://tcga-data.nci.nih.gov/tcga/) that include over 30 human cancer types (e.g. brain cancer, lung cancer, breast cancer, etc.) with multiple genomic profiles based on recent high-throughput platforms (e.g. RNA sequencing, single nucleotide polymorphisms, etc.). The TCGA program encouraged worldwide scientists to conduct comprehensive analyses of the large-scale dataset collected by the project, which contributes to the common goal of improving cancer diagnosis, treatment, and prevention (Chin et al., 2011; Tomczak, Czerwinska, & Wiznerowicz, 2015). Thus, the TCGA data are widely used and support meaningful scientific discoveries. However, it is challenging to track the TCGA data usage due to the data complexity and the lack of data identity. Researchers can use the TCGA via freely combining data from multiple types of cancer that were tested using multiple high-throughput platforms. In this preliminary study, we intended to identify the TCGA data citations by analyzing full-text literature mining.

2 Related Work

Text-mining methods have been developed to identify database citations from the literature by characterizing database entry accession numbers. Neveol et al. developed a machine learning method to extract data deposition statements from full-text literature (Neveol et al., 2011). Furthermore, they analyzed link curation between disposition databases (e.g. GEO and PDB) and the literature and proposed that text-mining tools can improve the links between literature and biological databases (Neveol et al., 2012). Kafkas, Kim, and McEntyre applied the patterns of ENA, Uniprot, and PDB accession numbers to identify database citations from full-text literature (Kafkas, Kim, & McEntyre, 2013) and from article supplemental files (Kafkas et al., 2015). Piwowar et al. investigated citation relationships between microarray databases (e.g. GEO and ArrayExpress) and the literature (Piwowar & Chapman, 2010; Piwowar & Vision, 2013). Yu et al. constructed a database link network from a set of pairs of databases that were co-mentioned in the methodology sections of full-text literature to track the database usage, connection, and evolution (Yu et al., 2015). These efforts have improved the understanding of data citations for specific databases that have identical accession numbers. However, few studies have been conducted to identify data-literature citation relationships for data generated by a scientific project in which the data is required to be shared but lacks a redefined identifier. In this study, we selected a publicly funded project and publicly available project data: TCGA (Tomczak, Czerwinska, & Wiznerowicz, 2015).

3 Methodology

To identify TCGA data usage from full-text articles, we proposed a computational framework (Figure 1). We collected TCGA-related full-text articles from PubMed Central, constructed a benchmark dataset which truly used the TCGA data, and analyzed data usage according to the specific cancer type and high-throughput platform.

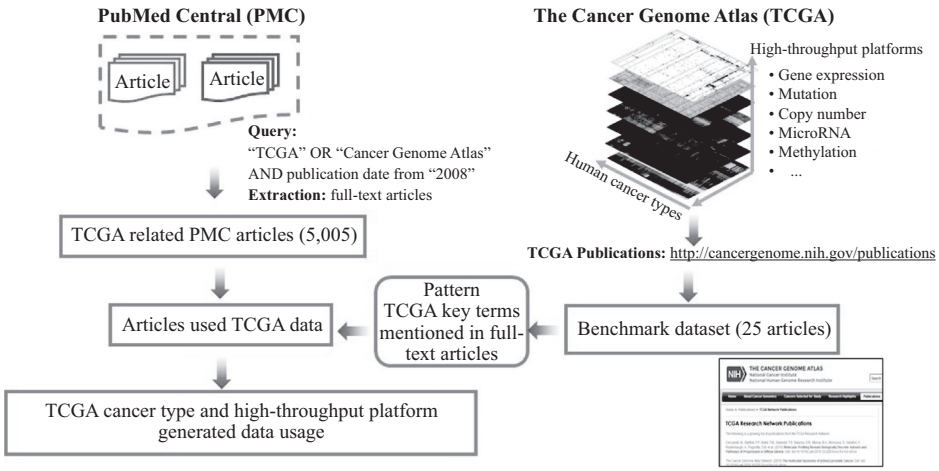

Figure 1. Computational workflow for identifying TCGA data usage.

3.1 TCGA-related Literature

PubMed Central (PMC, http://www.ncbi.nlm.nih.gov/pmc/), a publicly available literature archive, was used as the full-text article resource. In the literature contexts, the TCGA-related terms were mentioned in the form of both abbreviations and full-name descriptions. We extracted a set of full-text articles from PMC using the query "TCGA" or "Cancer Genome Atlas" and using the publication date of 2008 (Search term: (tcga OR "cancer genome atlas") AND ("2008" [Publication Date]: "2015" [Publication Date])). In total, 5,372 papers in XML format were collected as of October, 2015. Further, we removed the articles that merely mentioned "TCGA" or "Cancer Genome Atlas" in the article reference section. Then, 5,005 full-text articles remained and were included in the raw dataset for further analysis.

3.2 Benchmark Dataset

We collected 25 open access publications that used TCGA data, as confirmed by the TCGA Network, from the official website (http://cancergenome.nih.gov/publications). These articles constitute the benchmark dataset for the following

analysis. We attempted to characterize the TCGA data usage article patterns by analyzing the benchmark dataset. The location of the key term, "Cancer Genome Atlas" or its abbreviations, in the full-text literature was primarily investigated. Considering the varying structural composition of different journal articles, we divided the articles in six sections including title, abstract, introduction/background, method/material, results, and discussion/conclusion. We manually intervened when the PMC XML parser failed to identify the above sections.

3.3 TCGA Data Usage Analysis

We developed a full-text extraction method to parse the full-text articles in XML format, extracted metadata such as publication date and author country, and identified TCGA-related key terms as provided.

The cancer type and high-throughput platform are two characteristic classes of key words in the TCGA data usage statements. Here, the cancer type refers to a list of cancers investigated in the TCGA program, whereas the high-throughput platform refers to a list of high-throughput biotechnologies used by the TGCA investigators to test the cancer genomic information. In the TCGA program from 2005 to 2014, over 30 cancers were studied using microarray and next-generation sequencing platforms, consequently producing large-scale data, such as gene expression, exon expression, miRNA, copy number variation (CNV), single nucleotide polymorphism (SNP), loss of heterozygosity (LOH), mutations, DNA methylation, and protein expression. Referring to Disease Ontology (Kibbe et al., 2015) and the TCGA data matrix (TCGA data matrix, 2015), we developed a controlled vocabulary for the TCGA cancer type (Table 1) and high-throughput platform (Table 2).

Table 1. Examples of TCGA cancer-type concepts.

Concept ID	Name	TCGA defined terms [abbr] – [full name]	Synonyms	DO mapping
D0001	Glioblastoma	GBM – Glioblastoma Multiforme	Glioblastoma, GBM, adult glioblastoma multiforme, primary glioblastoma multiforme, spongioblastoma multiforme	DOID: 3068
D0002	Breast cancer	BRCA – Breast Invasive Carcinoma	Breast cancer, breast tumor, breast neoplasm, mammary cancer, mammary tumor, mammary neoplasm, malignant tumor of breast,	DOID: 1612
D0003	Ovarian cancer	OV – Ovarian Serous Cystadenocarcinoma	Ovarian cancer, ovarian tumor, ovarian neoplasm, ovary cancer, ovary tumor, ovary neoplasm, malignant tumor of ovary	DOID: 2394
D0004	Acute myeloid leukemia	LAML – Acute Myeloid Leukemia	Acute myeloid leukemia, AML, acute myeloblastic leukemia, acute myelogenous leukemia	DOID: 9119

Table 2. Examples of TCGA high-throughput platform concepts.

Concept ID	Name	TCGA-defined terms	Generated data
P0001	RNASeq	IlluminaGA_RNASeq, IlluminaHiSeq_RNASeq	Nucleotide sequence, gene expression
P0002	miRNASeq	IlluminaGA_miRNASeq	miRNAs, microRNA, microRNA sequence
P0003	SNP	Genome_Wide_SNP	SNPs, single nucleotide polymorphisms, CNV, copy number variation
P0004	Methylation	Human methylation	DNA methylation

As shown in Tables 1 and 2, the TCGA-defined terms were used to standardize the program-generated data description; however, they are not the terms used in the full-text articles. For example, in the results section of one article (PMCID: PMC3910500), it described the genomic landscape of glioblastoma using the whole-exome (WES), whole-genome sequencing (WGS), and RNA-Sequencing (RNA) (Brennan et al., 2013). To identify the TCGA cancer type and high-throughput platform concept from the free texts, we developed a named entity recognition method that is based on a biomedical text mining tool (Leaman, Islamaj, & Lu, 2013).

4 Results

4.1 Overview of the TCGA-related Publications

The number of TCGA-related articles increases as the program continues. Figure 2 shows the number of PMC articles related to the TCGA-related articles published from 2008 to 2015, and there were over 1,600 TCGA articles published in 2014. The 2015 reduction is due to data incompleteness as of September, 2015.

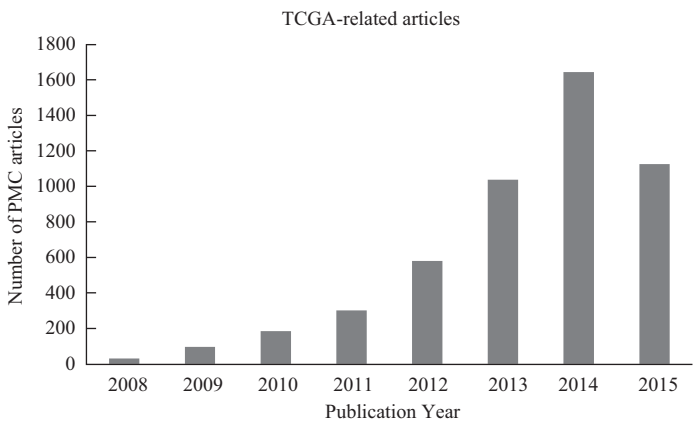

Figure 2. Number of TCGA-related publications in PMC.

TCGA data accumulation and data sharing contributed to the significant increase in TCGA publications. Phase I of the TCGA program (a 3-year pilot study) aimed to collect cancer tissues, process the biospecimen, apply high-throughput platforms to identify cancer genomic information, and analyze genetic changes involved in the cancer. Since 2009 (phase II), the data that were generated by the TCGA program have been centrally managed at the TCGA Coordinating Center and entered into public databases, allowing scientists to continually search, download, and analyze the data.

Figure 3 shows the geographical distribution of the TCGA-related publications. Researchers from 37 countries used the TCGA data in their studies, and the United States was the most productive one, followed by China, Canada, Australia, and Germany, etc.

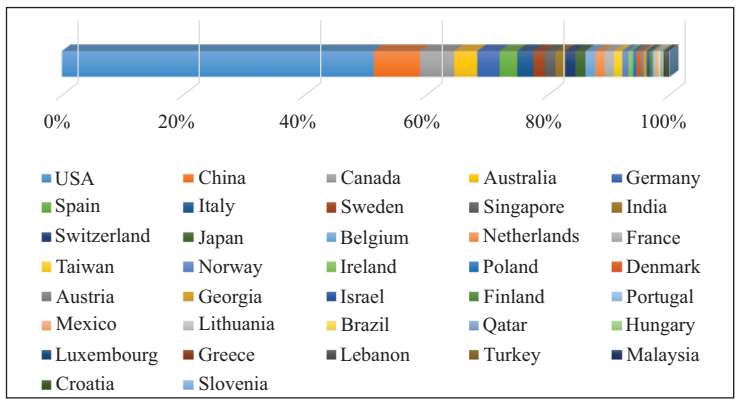

Figure 3. Geographical distribution of TCGA-related publications.

4.2 TCGA Key Terms That Were Mentioned in the Full-text Articles

We compared the TCGA key term features, TCGA term positions, and the TCGA-related concepts mentioned in the retrieved PMC articles (Section 3.1) and in the benchmark dataset (Section 3.2). Table 3 shows the true positive rate (TPR) of full-text articles in each dataset that have the TCGA key term features. The TCGA term (i.e. 'TCGA' or 'Cancer Genome Atlas') was mostly likely to appear in the results section in both the retrieved PMC article set (74%) and in the benchmark dataset (96%). Additionally, studies using the TCGA data are likely to describe the cancer type and high-throughput platform in the full-text articles of both datasets. Although there was a similar TCGA feature distribution within the retrieved PMC article set and within the benchmark dataset (χ^2 test, $p<0.05$), the article proportions were lower in the retrieved PMC article set than in the benchmark dataset. This is because

some articles in the retrieved set merely mentioned the TCGA term rather than actually using the data. In the following analysis, we focused on the PMC articles in which the TCGA term occurred in the methods/materials or the results section, which were the studies that were more likely to use TCGA data.

Table 3. Distribution of TCGA key terms in full-text articles.

Feature		Retrieved PMC article set (%)	Benchmark dataset (%)
TCGA term positon	Title	1	4
	Abstract	11	28
	Introduction/Background	12	20
	Method/Material	31	68
	Result	74	96
	Discussion/Conclusion	20	36
TCGA related concept mention	Cancer type mention	73	100
	Platform mention	66	96

4.3 TCGA Cancer Type and High-throughput Platform Generated Data Usage

To investigate the specific TCGA data usage, we identified the TCGA cancer type that was mentioned and the high-throughput platform that was mentioned in the methods/materials and in the results sections of the PMC full-text articles (Section 3.3). Figure 4 shows the proportion of different TCGA cancer types in the retrieved PMC article set. Glioblastoma (28%), lung cancer (18%), and breast cancer (11%) were the most frequent cancer types in which the data were used. Glioblastoma was the first cancer studied by the TCGA program, leading to TCGA infrastructure development that included data collection and sharing (Cancer Genome Atlas Research Network, 2008). Thus, this may be the major reason that the TCGA glioblastoma data were more frequently used.

As shown in Figure 5, the data generated by the RNASeq platform are the most widely used (48%). Compared with traditional DNA sequencing technology, RNA sequencing can help understand the transcriptome via precisely and rapidly deriving wide-range strand information, such as transcripts, isoforms, gene fusions, and non-coding RNAs (Wang et al., 2009). The TCGA data that is generated by the RNASeq platform provides researchers with standardized and comprehensive cancer transcriptome profiles to discover biomarkers related to tumorigenesis and metastasis (Peng et al., 2015).

5 Discussion

In this preliminary study, we conducted an investigation to track the use of scientific data that were generated by long-term government-funded program. We

Research Paper

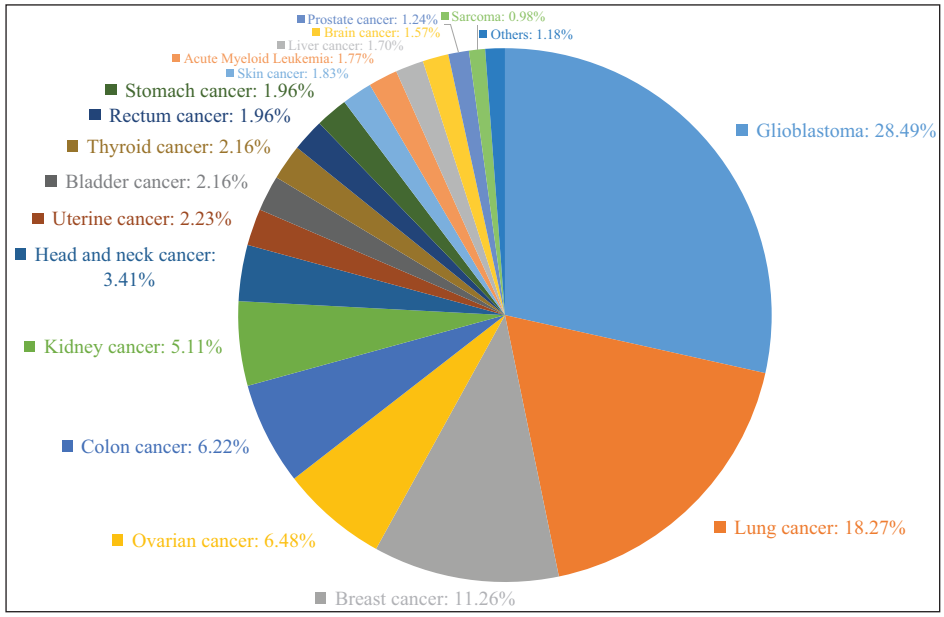

Figure 4. Distribution of TCGA cancer types.

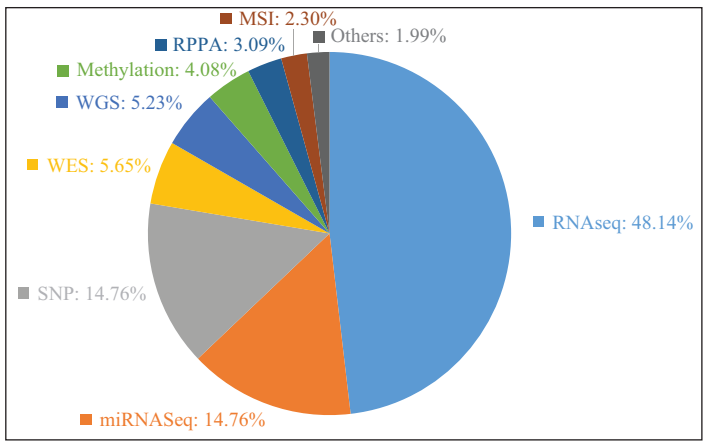

Figure 5. Distribution of the TCGA high-throughput platform.

selected the TCGA program and analyzed over 5,000 full-text articles that were collected from PMC. We constructed a benchmark dataset that truly used TCGA data, and we compared it with full-text articles retrieved from PMC. Furthermore, we built up a controlled vocabulary that was tailored for the TCGA program that describes the cancer type and high-throughput platform. Thus, it provides insights

into which specific data were used. Our work can contribute to scientific data and scientific literature integration. As shown in the box in Figure 6, the TCGA funding agencies manually collected the articles and linked the articles to their source data (TCGA publication, 2016). Our efforts may help develop an automatic method to identify recent publications that use TCGA data.

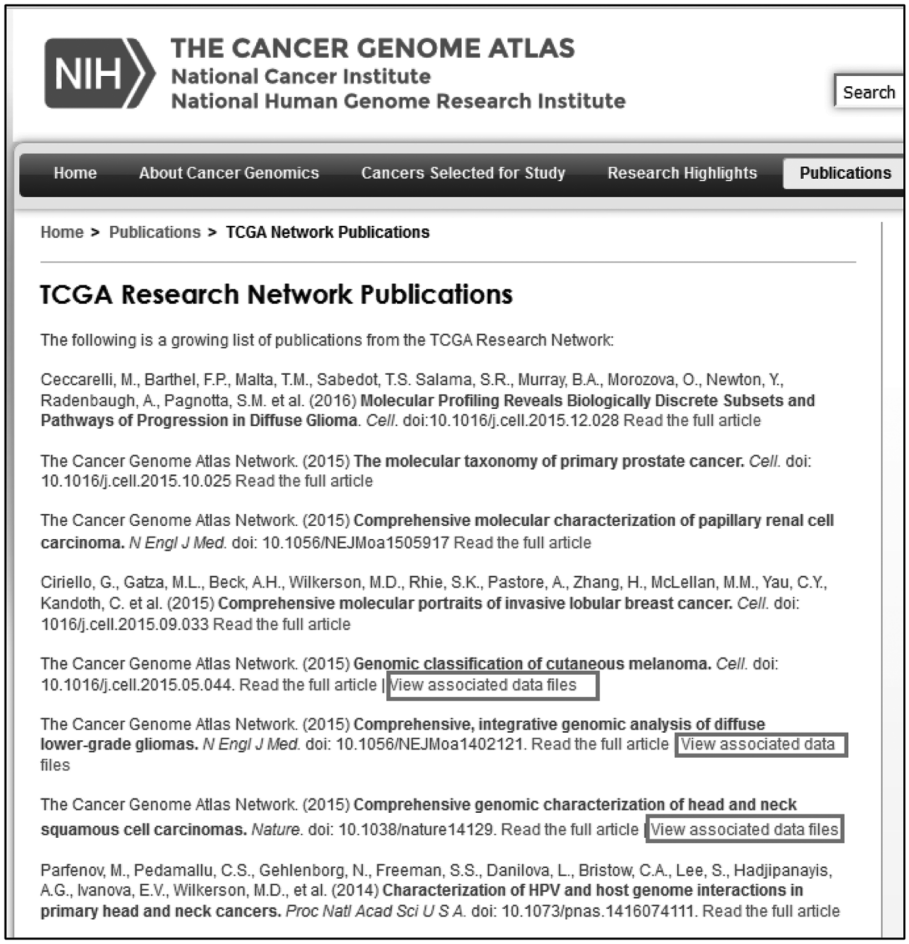

Figure 6. Manual link literature that includes TCGA data.

However, this study has limitations. (1) The benchmark set may cause a bias. We only collected 25 articles from the TCGA website to construct the benchmark dataset. The patterns of full-text articles that actually cite the TCGA data were not validated in a large scale dataset. Here, we only compared the TCGA term position and TCGA-related concept that were mentioned in the retrieved PMC articles and

in the benchmark dataset. In the future, we may manually construct a benchmark dataset that includes more full-text articles that actually cite TCGA data. (2) The identification performance of TCGA-related term requires evaluation. Here, we applied a biomedical text mining tool to identify the mentioned TCGA cancer type and high-throughput platform without validating the named entity recognition. (3) Natural language processing technology needs to confirm the relationships between cancer type and platform. The data usage statement in full-text literature describes which cancer type samples are tested by which platforms, however, we have not yet considered these specific relationships.

6 Conclusion

We present a workflow to identify scientific project-generated data citation via full-text article analysis, and we applied this workflow to track TCGA data citations via PMC literature analysis. In contrast to previous studies, the scientific data entries in our studies lacked predefined accession numbers. Although our preliminary study has limitations, this work is a step towards integrating literature with scientific data that are generated by a government-funded project. In future work, we expect to improve the construction of the scientific data citation benchmark dataset, normalize the full-text article sections, map the project self-defined vocabulary, and evaluate the performance of data citation identification.

Acknowledgements

This work was supported by the National Population and Health Scientific Data Sharing Program of China, the Knowledge Centre for Engineering Sciences and Technology (Medical Centre), and the Fundamental Research Funds for the Central Universities (Grant No.: 13R0101). The authors thank Yang Pan for the data processing during his summer internship.

Author Contributions

J. Li (li.jiao@imicams.ac.cn) conducted the study, wrote the manuscript, and revised the paper. S. Zheng (zheng.si@imicams.ac.cn) worked on the result analysis and drafted the manuscript. H. Kang (kang.hongyu@imicams.ac.cn) and Z. Hou (hou.zhen@imicams.ac.cn) collected data and developed the text-mining method. Q. Qian (qian.qing@imicams.ac.cn) proposed the research framework, discussed the results, and finalized the manuscript.

References

Bourne, P.E., Lorsch, J.R., & Green, E.D. (2015). Perspective: Sustaining the big-data ecosystem. Nature, 527(7576), S16–17.

Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R. (2013). The somatic genomic landscape of glioblastoma. Cell, 155(2), 462–477.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455(7216), 1061–1068.

Chin, L., Hahn, W.C., Getz, G., & Meyerson, M. (2011). Making sense of cancer genomic data. Genes & Development, 25(6), 534–555.

Green, E.D., Watson, J.D., & Collins, F.S. (2015). Human Genome Project: Twenty-five years of big biology. Nature, 526(7571), 29–31.

Kafkas, S., Kim, J.H., & McEntyre, J.R. (2013). Database citation in full text biomedical articles. PLoS One, 8(5), e63184.

Kafkas, S., Kim, J.H., Pi, X., & McEntyre, J.R. (2015). Database citation in supplementary data linked to Europe PubMed Central full text biomedical articles. Journal of Biomedical Semantics, 6, 1.

Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., Parkinson, H., & Schriml, L.M. (2015). Disease Ontology 2015 Update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Research, 43(Database issue), D1071–1078.

Leaman, R., Islamaj, D.R., & Lu, Z. (2013). DNorm: Disease name normalization with pairwise learning to rank. Bioinformatics, 29(22), 2909–2917.

National Science Board (2005). Long-lived digital data collections: Enabling research and education in the 21st century. Retrieved on Oct 20, 2015, from http://www.nsf.gov/pubs/2005/nsb0540/

Neveol, A., Wilbur, W.J., & Lu, Z. (2011) Extraction of data deposition statements from the literature: A method for automatically tracking research results. Bioinformatics, 27, 3306–3312.

Neveol, A., Wilbur, W.J., & Lu, Z. (2012). Improving links between literature and biological data with text mining: A case study with GEO, PDB and MEDLINE. Database (Oxford), 2012, bas026.

Peng, L., Bian, X.W., Li, D.K., Xu, C., Wang, G.M., Xia, Q.Y., & Xiong, Q. (2015). Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. Scientific Report, 5, 13413.

Piwowar, H., & Chapman, W. (2010). Recall and bias of retrieving gene expression microarray datasets through PubMed identifiers. Journal of Biomedical Discovery and Collaboration, 5, 7–20.

Piwowar, H., & Vision, T.J. (2013). Data reuse and the open data citation advantage. Peer J, 1, e175.

TCGA Data Matrix (2015). Retrieved on Oct. 20, 2015, from https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm

TCGA Publications (2016). Retrieved on Jan. 28, 2016, from http://cancergenome.nih.gov/publications.

Tomczak, K., Czerwinska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Contemporary Oncology, 19(1A), A68–77.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57–63.

Yu, Q., Ding, Y., Song, M., Song, S., Liu, J., & Zhang, B. (2015). Tracing database usage: Dectecting main paths in database link network, Journal of Informetrics, 9(1), 1–15.

Research Paper