Using Semantic Web Technologies to Enable Cancer Genomics Discovery at Petabyte Scale

Cancer Informatics
Volume 17: 1–7

DOI: 10.1177/1176935118774787

Jovan Cejovic, Jelena Radenkovic, Vladimir Mladenovic, Adam Stanojevic, Milica Miletic, Stevan Radanovic, Dragan Bajcic, Dragan Djordjevic, Filip Jelic, Milos Nesic, Jessica Lau, Patrick Grady, Nick Groves-Kirkby, Deniz Kural and Brandi Davis-Dusenbery

Seven Bridges Genomics Inc., Cambridge, MA, USA.

ABSTRACT: Increased efforts in cancer genomics research and bioinformatics are producing tremendous amounts of data. These data are diverse in origin, format, and content. As the amount of available sequencing data increase, technologies that make them discoverable and usable are critically needed. In response, we have developed a Semantic Web–based Data Browser, a tool allowing users to visually build and execute ontology-driven queries. This approach simplifies access to available data and improves the process of using them in analyses on the Seven Bridges Cancer Genomics Cloud (CGC; www.cancergenomicscloud.org). The Data Browser makes large data sets easily explorable and simplifies the retrieval of specific data of interest. Although initially implemented on top of The Cancer Genome Atlas (TCGA) data set, the Data Browser's architecture allows for seamless integration of other data sets. By deploying it on the CGC, we have enabled remote researchers to access data and perform collaborative investigations.

KEYWORDS: TCGA, cancer, genomics, cloud, Semantic Web

RECEIVED: March 17, 2017. **ACCEPTED:** August 3, 2017.

TYPE: Sequencing for the Masses: Another Desktop Revolution - Review

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under contract no. HHSN261201400008C.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Brandi Davis-Dusenbery, Seven Bridges Genomics Inc., 1 Main Street, Suite 500, Cambridge, MA 02142, USA.
Email: brandi@sbgenomics.com

Introduction

Effective access to large-scale genomic data and associated clinical data is invaluable for accelerating research in cancer prevention, diagnosis, and treatment. As the amount of available sequencing data increases technologies to make this data findable, discoverable, and usable are critically needed. Here, we present our approach to democratizing access to cancer genomics data through the Seven Bridges Cancer Genomics Cloud (CGC; www.cancergenomicscloud.org), which includes a Semantic Web–based Data Browser that enables researchers to visually query large, complex data sets.

The CGC is a US National Cancer Institute pilot designed to democratize access and analysis of massive cancer genomics data sets including The Cancer Genome Atlas (TCGA).[1] TCGA comprises samples from 33 tumor types taken from over 11 000 patients. At over 1 petabyte in size, TCGA is one of the largest and most complete genomic data sets in the world. TCGA combines comprehensive clinical information with genomic analyses including whole genome, whole exome, RNA, and microRNA sequencing, as well as methylation and protein analyses.

Previous and current TCGA repositories use form-based browsers,[2–4] enabling users to identify and select subsets of the TCGA data set by filtering the data based on values of one or more fields specified by their respective metadata schemas. This approach, although easy to implement, lacks the functionalities needed for executing complex data queries without additional browsing. Furthermore, researchers have to download and manually parse metadata, then use them to select data of interest. Due to the size and complex overall structure of TCGA, it is challenging for researchers to identify and retrieve relevant files within the data set. There is a clear need for an improved method of querying TCGA that would enable researchers to effectively access the information within.

The GDC Data Portal[5] is a more advanced example of this class of tools. While offering rich filtering capabilities and basic statistics, it focuses primarily on cases and files. Researchers are unable to create queries based on more complicated entity relationships. There is a clear need for an improved method of querying cancer genome data in general.[6] More specifically, the size and complex overall structure of TCGA and other consortia-developed data sets make it challenging for researchers to effectively access the information within. Moreover, search across multiple diverse data sets, potentially with divergent ontologies, remains particularly challenging.

To address inherent problems in linking information within TCGA in TCGA's complex structure and provide a solid foundation and infrastructure for integrated querying with other data sets,[7] we developed a Semantic Web–based solution for data querying within the CGC. We have built a rich knowledge base (see Box 1) containing more than 150 clinical, biospecimen, and analytical properties that describe cancer genomics data. Our solution achieves a flexible data model that simplifies expansion, reuse, and iterative revision. Furthermore, to ensure that our solution's functionality is widely accessible to all researchers, we have

Box 1. Key definitions.

- Semantic Web: An extension of the World Wide Web where data are defined in a meaningful way, using standardized formats that facilitate exchange, processing, and integration.
- Entity: Generally, any abstract or concrete thing that fundamentally exists in the current domain. In our TCGA knowledge base, examples include patient information, clinical data such as whether a patient received radiation therapy, and file format.
- Ontology: A practical application of philosophical ontology—a formal naming and definition of the types, properties, and interrelationships of entities.
- Resource Description Framework (RDF): A specification of data representation in the form of triples, statements following the subject-predicate-object pattern. Subjects are entities, objects are traits of subjects, and predicates express relationships between subjects and objects.
- RDF Schema (RDFS): A set of classes and properties defined using RDF, used for basic ontology description.
- Web Ontology Language (OWL): A more expressive family of languages based on RDF, used to represent ontologies in conjunction with RDFS.
- SPARQL: A recursive acronym for SPARQL Protocol and RDF Query Language; a query language for RDF data manipulation and retrieval.
- Knowledge base: An organized repository of knowledge represented in triple format, on which semantic and logical operations are performed.

developed a Data Browser that enables interactive data exploration by visually building SPARQL queries.

The CGC hosts both open and controlled access cancer genomic data. While all researchers are able to freely register and use the CGC to analyze open access genomic data, or upload their private data, access to controlled data is governed by dbGaP Data Use Requests. In addition to supporting novel ways to discover data, the CGC provides a rapid path to reproducible and scalable analysis of genomic data via Rabix, our implementation of the Common Workflow Language. Workflows are executed on Amazon Web Services (AWS) and researchers pay only for the computational resources that are used. A standard RNA-Seq analysis costs US $1 to US $2 per sample and use of Spot instances can further reduce costs depending on the global AWS capacity. New users are able to access computational and storage credits as they are learning the system.

Since its public launch in February 2016, within two years, more than 3200 users have joined the CGC. Researchers have used CGC to quantify differential expression of mammary tumor-associated lncRNAs[8] to retrieve RNA-Seq data from TCGA[9] and to test cancer neoantigen caller workflows in the cloud.[10]

Findable, Discoverable, and Usable Cancer Data

Building a Semantic Web knowledge base

Through careful analysis of the structure of TCGA data and metadata, we identified a set of central concepts for efficient querying and filtering. These concepts were modeled with an ontology using the Semantic Web specifications RDF, RDFS, and OWL (Figure 1; Box 1). The ontology was engineered using the Protégé framework[11] and the Python library rdflib.[12] In accordance with this ontology, a knowledge base was created and populated with prepared information obtained from raw TCGA data.

Two different base classes of entities naturally emerged: Entity and Utility. The Entity class comprises entities with a central role in analysis and investigation such as subpatient information, clinical data (eg, radiation therapy, follow-up, new tumor event), and biospecimen data (eg, sample, portion, slide, analyte, aliquot). The Utility class includes entities that describe, explain, quantify, or categorize instances of the Entity class. For example, it can be used to define a patient's sex or disease type or to define a file's data format or sample type. Each of these entities is a set of well-known, community-recognized, enumerated values.

There are also 2 different types of properties: object and datatype properties. Relationships between entities are modeled with object properties (eg, when a File contains data for a Sample or when a Case has a Sample). Object properties are also used to associate instances from the 2 different classes (eg, each File has an associated Data Format, an instance of the Utility class). Datatype properties provide identifiers, labels, or physical values for entities. Datatype properties can include barcodes, file names, amounts, and concentrations.

To connect the TCGA knowledge base with the visual Data Browser, we created 2 additional ontologies: query service ontology and extract, transform, load (ETL) ontology. They both extend the base TCGA ontology and provide useful information for their respective tasks. The query service ontology provides additional information crucial for visual querying—flags denoting which entities could be used to start a query, groupings of properties into meaningful categories, and information necessary for integration with the rest of the CGC platform (such as physical location of the underlying TCGA files). The ETL ontology facilitates data import to RDF repositories used by the Data Browser. It precisely describes the structure of raw source data files (usually in XML, CSV, or other text format) so the ETL tools know which sections are of interest for the data set at hand. It is used by the ETL tool to inspect and record the properties and relationships between entities.

Populating the knowledge base

We imported TCGA data to our own cloud file system from multiple sources including the TCGA Data Portal,[4] CGHub,[2]

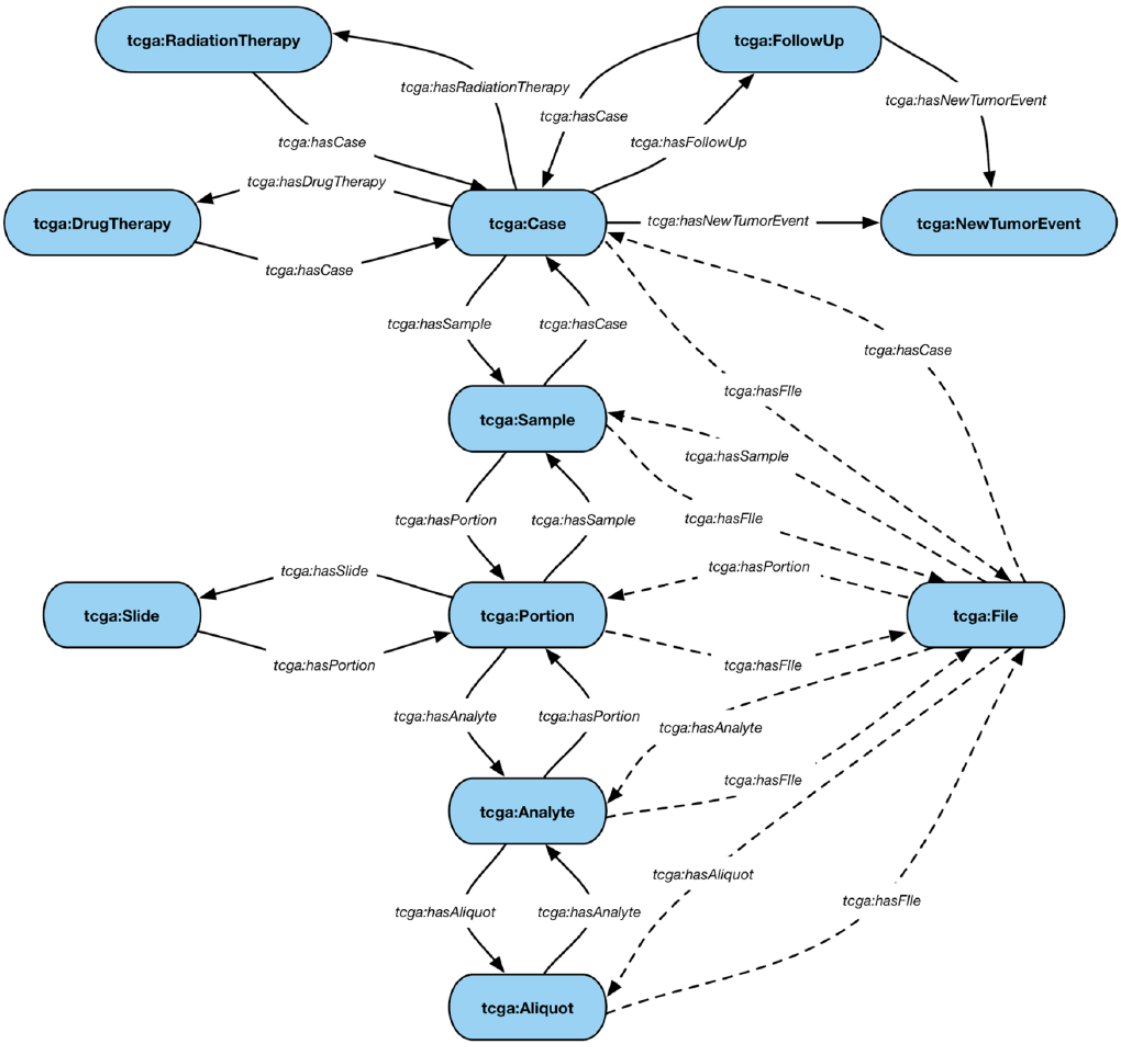

Figure 1. Subset of TCGA ontology. Relationships between Entity subclasses. Blue ovals represent subclasses of the Entity class, and arrows represent object properties that describe relationships between the subclasses. TCGA indicates The Cancer Genome Atlas.

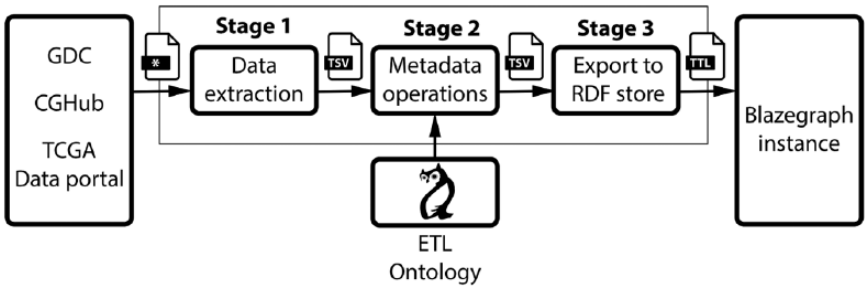

Figure 2. Populating the knowledge base using ETL ontology. TCGA data are imported from diverse sources, metadata are extracted and harmonized, and RDF data are created and exported to a Blazegraph database. ETL indicates extract, transform, load; RDF, Resource Description Framework; TCGA, The Cancer Genome Atlas.

and Genomic Data Commons.[2,3] We built a Python-based tool that relies on the ETL ontology to harmonize the multiple data formats from the different sources, extract relevant metadata from source files, and process the metadata to produce the RDF knowledge base.

Our ETL tool works in 3 stages (Figure 2). First, the ETL tool acquires files (in their original format), from our file system, extracts information of interest, and creates TSV files containing TCGA entity data. Second, the tool performs merging, adding, filtering, and transforming operations to produce TSV files that contain record entities and properties information. This is accomplished by consulting the ETL ontology. Third, RDF data are produced by generating triples for each row of each TSV file. The RDF data are then uploaded

to a Blazegraph database server,[8] which is RDF and SPARQL compliant, and/or stored in RDF files using the rdflib library.

The Blazegraph server delivers extracted data to the Data Browser from 2 separate RDF stores: the ontology store and the knowledge base. The ontology store is reserved for the query service ontology, whereas the knowledge base contains materialized data for all entities and relationships described within the ontology (16 127 033 triples). The majority represent relationships between TCGA domain entities and related files, file-related properties, and relationships between the domain entities themselves.

Architecture of the CGC Data Browser

To make building complex queries accessible to a wide audience, we developed a visual query engine—the CGC Data Browser (Figure 4). The Data Browser is accessed via a Web browser as part of the CGC platform (www.cancergenomicscloud.org). The CGC is open to all cancer researchers worldwide, who can create a free profile online or log in via their eRA Commons or NIH CIT account. The Data Browser makes it easy for researchers to quickly search TCGA across more than 100 different metadata properties to find and access data of interest.

A high level overview of the Data Browser architecture is presented in Figure 3. In brief, the working space is initialized when a researcher loads the Data Browser. There is a dedicated back-end service for the Data Browser—the query service—which encapsulates the logic for converting the visual representation into a SPARQL query, result retrieval, and formatting. The Data Browser's front end is implemented in JavaScript and relies on a proprietary SVG library to render the graphical query representation.[13]

The Data Browser relies on a starting configuration which is obtained from the query service ontology (which in turn contains all the information from the domain ontology, in this case, TCGA). The configuration is a result of a simple query requesting all the subclasses of the Entity and Utility classes and their corresponding properties. This configuration is used as a base to construct a JSON representation of the query which is sent to the query service back end for execution.

Visual querying with the CGC Data Browser

Queries are built by adding entities and filtering them by their properties and associated values (Figure 4A). In addition to selecting entities from scratch, there are a number of other ways for researchers to start building a query. First, the Data Browser provides example queries, which researchers can modify. Second, the Data Browser features a search box, which allows researchers to query data by Universally Unique Identifiers (UUID), TCGA Barcodes (ID), or file names (Figure 4B). Finally, the CGC Case Explorer[14] is a separate tool that allows for visual exploration of genotypic information within cancer subtypes. A dynamic scatterplot displays the

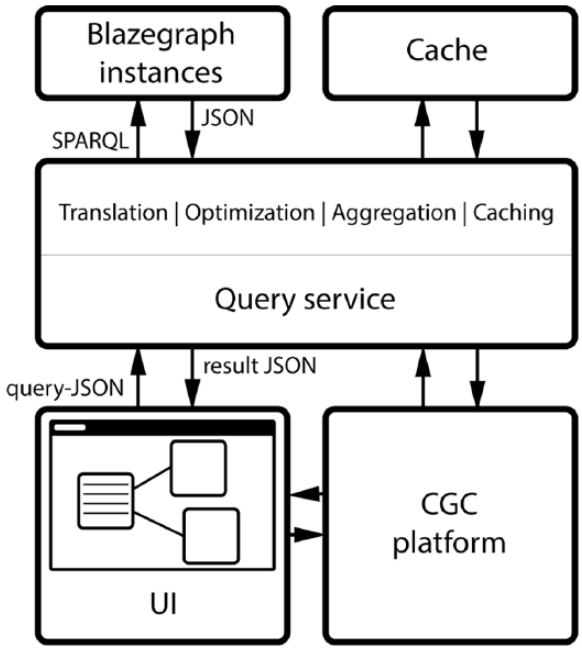

Figure 3. High-level architecture behind the Data Browser. The user constructs queries in the UI by connecting nodes corresponding to entities available from the starting configuration, obtained from the query service ontology. Each connection represents the appropriate object property from the ontology. A JSON protocol defined between the front end and the query service captures necessary information about the current query and its context. The front end serializes the graphical query representation into this protocol-defined query-JSON. The serialized query is then transferred to the back end. A custom SPARQL query builder then converts it to the SPARQL format. Next, the query is executed on the end points associated with the current data set. Finally, the results are cached. The resulting data is sent to the front end to be displayed on the graphical user interface. Users are then able to extract the relevant files and import them to one of their projects on the CGC platform. CGC indicates Cancer Genomics Cloud.

distribution of cases based on gene expression, mutation type, and copy number, allowing researchers to group cases based on genotypic parameters (Figure 5). After cases are selected in the Case Explorer, they can be brought to the Data Browser for further filtering. Through these mechanisms, the Data Browser enables researchers with diverse expertise to visually build complex queries and retrieve the resulting files for analysis.

As a query is built, the Data Browser displays counts that indicate how many instances match a query (Figure 4C). Query results are listed in a table (Figure 4D) and visualized in graphs that show the distributions of the instances of each entity (Figure 4E). After a query is completed, the resulting files can be immediately retrieved and analyzed. The files can be copied to a CGC project and analyzed with tools available on the platform (Figure 4F). Alternatively, the results table can be exported in various formats (Figure 4G). Furthermore, queries can be saved so the researchers can easily continue their work (Figure 4B).

To evaluate the Data Browser, we recreated a query from a recent study.[15] In this article, the researchers wanted to identify and access TCGA RNA-Seq gene expression data from a subset of patients, African American female patients who were

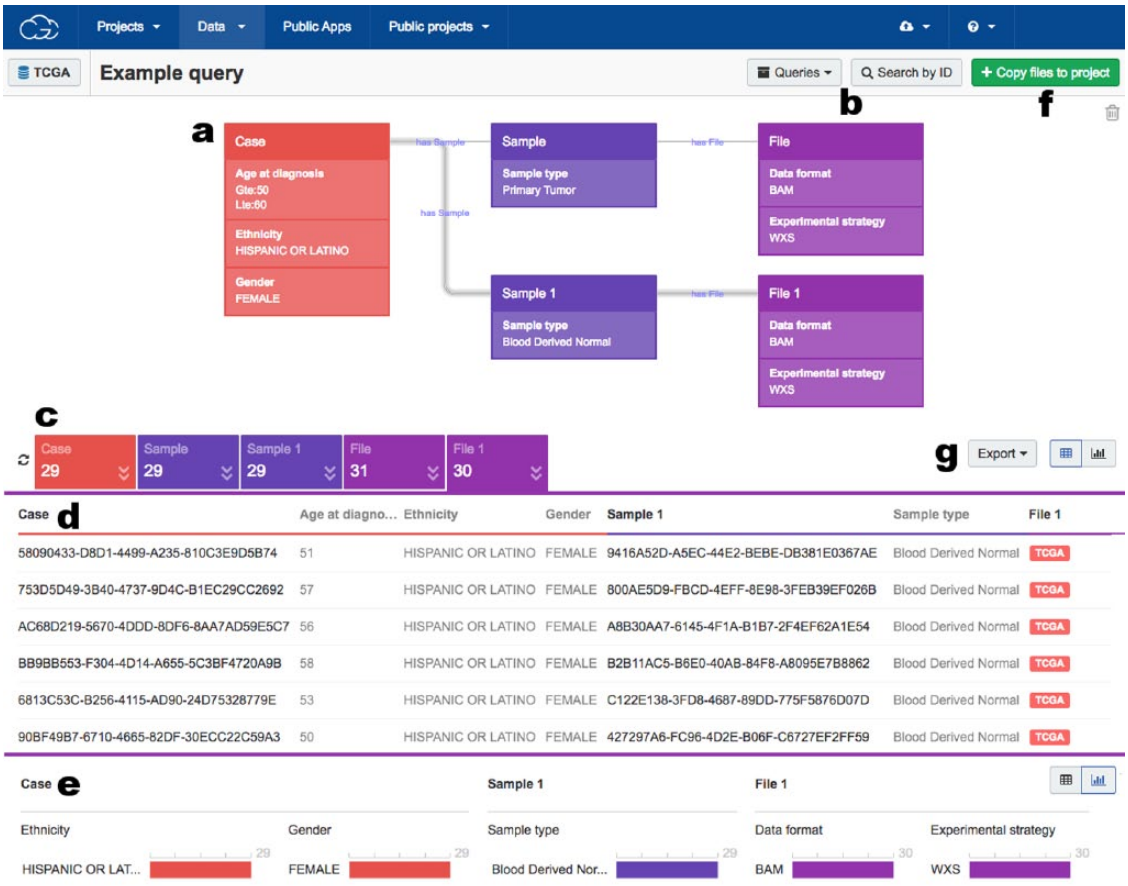

Figure 4. The CGC Data Browser. An example query (A) identifies all BAM files of tumor and normal samples from Hispanic or Latino female patients, diagnosed with invasive breast carcinoma between the ages of 50 and 60 years, analyzed using whole exome sequencing. For this query, the starting entity is Case. The entity is filtered by the property Age at diagnosis, with a value range of between 50 and 60 years. The entity is additionally filtered by the Ethnicity, Disease type, and Gender properties. After a starting entity is selected, a list of associated entities is displayed so that the query can be further refined. Here, the second entity selected is Sample, which is filtered by the property Sample type, with the values of Primary tumor and Blood-derived normal. The third entity, File, enables selection of BAM files from whole exome sequencing studies. In addition to selecting entities from scratch, the Data Browser enables queries based on examples and ID searches (B). During the query, counts (C) of how many instances match the query are displayed and can be refreshed. Query results are displayed in a table (D) or distribution graphs (E). Files identified by the query can be immediately accessed by copying them to a CGC project (F). Alternatively, the results table can be exported in CSV, JSON, or TSV formats (G). CGC indicates Cancer Genomics Cloud.

diagnosed with stage I, II, or III breast carcinoma between 1988 and 2013. The researchers had to download archives containing clinical metadata and gene expression data, then manually parse the metadata and use them to select the data of interest.

We performed the same search by building a query in our Data Browser. A query can be assembled rapidly, without downloading any data and without downloading any data; assembling the query took less than 30 seconds. The resulting gene expression files could be immediately accessed for further analysis on the CGC platform.

Programmatic interaction and connecting to other semantic databases

In addition to the Data Browser, the knowledge base can be queried programmatically using a simplified API.[16] The RESTful API can be used to browse and query TCGA and other data sets available on the CGC using a simplified JSON-based query language. It is offered as a simpler alternative to SPARQL but does not support its full feature set and is primarily used to integrate the CGC with other applications and automate the process of querying metadata. API requests are made over HTTP, and information is sent and received in JSON format.

The patterns from concepts from TCGA ontology can be generalized and applied to other data sets as well. Many data sets contain information which could be easily categorized into one of the 2 classes, Entity and Utility. As the Data Browser only relies on these 2 classes, it could work with any arbitrary ontology which follows this pattern. The ETL process is inherently flexible and can be configured to extract and harmonize metadata from other sources data sets, by modifying the ETL ontology to describe the format of new inputs. Currently, the Cancer Imaging Archive (TCIA),[17] Therapeutically Applicable Research to Generate Effective Treatments (TARGET),[18]

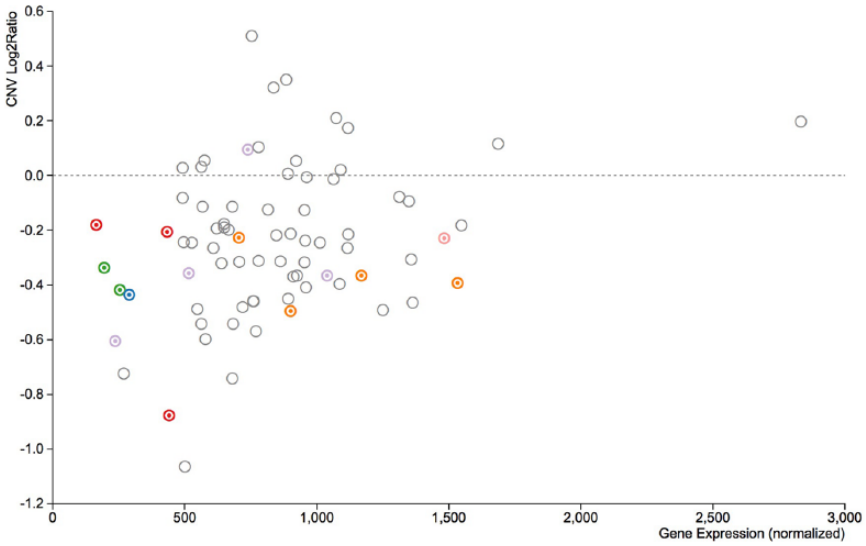

Figure 5. The CGC Case Explorer. Users can use the Case Explorer to select subsets of cases for further querying within the Data Browser. This example shows TP53 expression and copy number variation in patients with adenoid cystic carcinoma in TCGA. The color of each circle indicates TP53 mutation type: missense mutations (orange), frameshifts (red and blue), in-frame deletions (pink), nonsense mutations (green), and splice site mutations (lavender). Empty circles indicate cases where sequence is consistent with the reference.

Clinical Proteomic Tumor Analysis Consortium (CPTAC)[19] and the International Cancer Genome Consortium (ICGC)[20] data sets are available on the CGC, as well as the Cancer Cell Line Encyclopedia (CCLE)[21] data set, under the "legacy data sets" category.

Uptake and Future Directions

Since the early access release of the CGC in November 2015, we have been collecting feedback from researchers who use the Data Browser. We received positive feedback regarding several Data Browser features: researchers can build rich queries that would be difficult or impossible to create using alternative approaches, researchers can use branch points to find data that match multiple criteria, and researchers can see the number of entities that match a query. However, some aspects of the Data Browser could be improved: match counts need to be refreshed whenever the query changes, building complex queries can be time-consuming or difficult to interpret. We intend to improve subsequent versions of the Data Browser based on this feedback.

The promise of the Semantic Web is its ability to unify and query data from disparate sources. The RDF triple format is amenable to this task, but identifying similar concepts between different data sets and creating meaningful links between them can be a daunting task.[22] Future work will focus on enabling this integrated querying between data sets from similar domains. Currently, this can be achieved in 2 ways. Query rewriting is the process of translating a query corresponding to the concepts of the first ontology into a semantically similar query corresponding to the concepts of the second.[23] This is the focus of very active research in the Semantic Web community. The other approach is to extract shared concepts into a single, parent ontology and use it to query both data sets. We are currently investigating both approaches.

Other future work includes bringing more data sets to the CGC platform, providing query recommendations to improve the learning process, and investigating how to deploy the Data Browser on researchers' custom data.

Conclusions

As the amount of available sequencing data increase at a beyond-exponential rate, technologies that make data discoverable and usable will be necessary. We designed the CGC to democratize access to massive cancer genomics data sets containing petabytes of information, starting with TCGA. To address inherent problems in linking information in TCGA's complex structure and future scalability issues in relation to other data sets, we developed a Semantic Web–based solution for data querying in the CGC. This knowledge base contains more than 150 clinical, biospecimen, and analytical properties that describe cancer genomics data. To further streamline data access, we developed a Data Browser that enables interactive data exploration.

By employing a Semantic Web–based solution for data querying in the CGC, we provide connectivity and scalability not only within TCGA but within other semantic databases as well. The ETL is inherently flexible to extract and harmonize metadata from other data sets, allowing for integrated analyses. Our flexible data model simplifies expansion, reuse, and iterative revision.

Proliferation of RDF data, especially in the life sciences, led us to deploy the Semantic Web as a solution for the CGC. Not only does this approach enable powerful ways of exploring and using TCGA data, the extensibility of this approach will also be useful in enabling researchers to learn from other large,

heterogeneous data sets including pediatric cancer genomics efforts[24] and precision medicine initiatives such as the Million Veteran Program.[25]

Acknowledgements

The authors thank the whole Seven Bridges team for helpful discussions, development, and feedback. The authors are grateful to the patients who donated their samples to TCGA and the many researchers who have generated this data.

Author Contributions

The Data Browser and the related tools were developed by all authors jointly. JC has devised the initial architecture of the query service and implemented it together with AS, MM, SR and VM. The data model and the ontologies were developed by VM, AS, JR and MN. The ETL tool was developed by AS and MM. The front end work was split between DD, DB and FJ. BD-D and DK oversaw the project and guided the integration with the rest of the CGC platform. JL, PG, and NG-K supported development of article and the manuscript was drafted by JC, AS, VM, JR, JL, PG, and NG-K with critical revisions provided by BD-D and JC. All authors reviewed the manuscript.

Disclosures and Ethics

As a requirement of publication, authors have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality, and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

REFERENCES

1. The future of cancer genomics. *Nat Med*. 2015;21:99.
2. Wilks C, Cline MS, Weiler E, et al. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database*. 2014;2014:bau093. doi:10.1093/database/bau093.
3. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl J Med*. 2016;375:1109-1112.
4. The Cancer Genome Atlas. Data portal. https://tcga-data.nci.nih.gov/docs/publications/tcga/. Accessed May 17, 2017.
5. GDC data portal. https://portal.gdc.cancer.gov. Accessed May 17, 2017.
6. Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. *Genes Dev*. 2011;25:534-555.
7. Chen H, Yu T, Chen JY. Semantic Web meets integrative biology: a survey. *Brief Bioinform*. 2013;14:109-125.
8. Diermeier SD, Chang KC, Freier SM, et al. Mammary tumor-associated RNAs impact tumor cell proliferation, invasion, and migration. *Cell Rep*. 2016;17:261-274.
9. Collado-Torres L, Nellore A, Kammers K, et al. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol*. 2017;35:319-321.
10. Bais P, Namburi S, Gatti DM, Zhang X, Chuang JH. CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics*. 2017;33:3110-3112. doi:10.1093/bioinformatics/btx375.
11. Musen MA; Protégé Team. The Protégé project: a look back and a look forward. *AI Matters*. 2015;1:4-12.
12. RDFLib. RDFLib/rdflib. GitHub. https://github.com/RDFLib/rdflib. Accessed May 17, 2017.
13. flipio. flipio/GraphBuilder. GitHub. https://github.com/flipio/GraphBuilder. Accessed May 17, 2017.
14. The CGC Knowledge Center. About the case explorer. http://docs.cancergenomicscloud.org/docs/about-the-case-explorer. Accessed May 17, 2017.
15. Keenan T, Moy B, Mroz EA, et al. Comparison of the genomic landscape between primary breast cancer in African American versus white women and the Association of Racial Differences with Tumor Recurrence. *J Clin Oncol*. 2015;33:3621-3627.
16. The CGC Knowledge Center. About the datasets API. http://docs.cancergenomicscloud.org/docs/about-the-datasets-api. Accessed May 17, 2017.
17. The Cancer Imaging Archive (TCIA) - helping to connect phenotypes to genotypes. http://www.cancerimagingarchive.net/. Accessed June 4, 2018.
18. The Therapeutically Applicable Research to Generate Effective Treatments (TARGET). https://ocg.cancer.gov/programs/target/overview/. Accessed June 4, 2018.
19. Clinical Proteomic Tumor Analysis Consortium (CPTAC) https://proteomics.cancer.gov/programs/cptac/. Accessed June 4, 2018.
20. International Cancer Genome Consortium (ICGC) https://icgc.org/. Accessed June 4, 2018.
21. Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483:603-607.
22. Pasquier C. Biological data integration using Semantic Web technologies. *Biochimie*. 2008;90:584-594.
23. Correndo G, Salvadores M, Millard I, Glaser H, Shadbolt N. SPARQL query rewriting for implementing data integration over linked data. Paper presented at: 1st International Workshop on Data Semantics (DataSem 2010); March 22, 2010; Lausanne.
24. Cavatica democratizing access to pediatric genomics data. http://www.cavatica.org/. Accessed May 17, 2017.
25. Gaziano JM, Concato J, Brophy M, et al. Million veteran program: a megabiobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214–223. doi: 10.1016/j.jclinepi.2015.09.016.