Linked Cancer Genome Atlas Database*

Muhammad Saleem
Universität Leipzig, IFI/AKSW,
PO 100920, D-04009 Leipzig
saleem@informatik.uni-leipzig.de

Shanmukha S. Padmanabhuni
Digital Enterprise Research Institute, National University of Ireland Galway (NUIG), Ireland
shanmukha.sampath@deri.org

Axel-Cyrille Ngonga Ngomo
Universität Leipzig, IFI/AKSW,
PO 100920, D-04009 Leipzig
ngonga@informatik.uni-leipzig.de

Jonas S. Almeida
Div. Informatics, Dept. of Pathology, University of Alabama, Birmingham
jalmeida@uab.edu

Stefan Decker
Digital Enterprise Research Institute, National University of Ireland Galway (NUIG), Ireland
stefan.decker@deri.org

Helena F. Deus
Foundation Medicine Inc. One Kendal Square Cambridge, MA
hdeus@foundationmedicine.com

ABSTRACT

The Cancer Genome Atlas (TCGA) is a multidisciplinary, multi-institutional pilot project to create an atlas of genetic mutations responsible for cancer. One of the aims of this project is to develop an infrastructure for making the cancer related data publicly accessible, to enable cancer researchers anywhere around the world to make and validate important discoveries. However, data in the cancer genome atlas[1] are organized as text archives in a set of directories. Devising bioinformatics applications to analyse such data is still challenging, as it requires downloading very large archives and parsing the relevant text files in order to collect the critical co-variates necessary for analysis. Furthermore, the various types of experimental results are not connected biologically, i.e. in order to truly exploit the data in the genome-wide context in which the TCGA project was devised, the data needs to be converted into a structured representation and made publicly available for remote querying and virtual integration. In this work, we address these issues by RDFizing data from TCGA and linking its elements to the Linked Open Data (LOD) Cloud. The outcome is the largest LOD data source (to the best of our knowledge) comprising of over 30 billion triples. This data source can be exploited through publicly available SPARQL endpoints, thus providing an easy-to-use, time-efficient, and scalable solution to accessing the Cancer Genome Atlas. We also describe showcases which are enabled by the new linked data representation of the Cancer Genome Atlas presented in this paper.

Categories and Subject Descriptors

H.3.5 [**Information storage and retrieval**]: On-line Information Services—*data sharing, web-based service*; H.2.4 [**Database Management**]: Systems—*Distributed databases, query processing*

Keywords

TCGA, SPARQL, LOD

1. INTRODUCTION

The Cancer Genome Atlas (TCGA)[2] is an effort led by the National Cancer Institute[3] and aims to characterize and sequence 33 cancer types from 9000 patients at the molecular level. The ultimate goal of the project is to collect and make publicly available the data necessary to produce an Atlas of the genomic alterations responsible for the initiation and progression of cancer. TCGA offers data categorized into three data levels: raw data (level 1), normalized data (level 2) and processed data (level 3). To date, a total of 21 types of data have been collected for each patient, making up a total of 147,645 raw data files, of which 53,694 contain level 3 (processed) data, summing up to a total of 12.7 terabytes of data. According to information in the TCGA portal, this is only 46% of the expected data with new data being submitted every day. In this paper, only level 3 data is of interest as it is the data upon which analytics is performed.

[1] https://tcga-data.nci.nih.gov/tcga/

[2] https://tcga-data.nci.nih.gov/tcga/

[3] http://www.cancer.gov

TCGA is a valuable resource for hypothesis-driven translational research as all of its data results from direct experimental evidence. Analysis of such evidence within cancer research has led in recent years to clinically relevant findings in the genetic mark-ups of different cancers and was at the forefront of a coordinated worldwide effort towards making more molecular results from cancer analysis publicly available [5]. Other big data cancer research initiatives such as the international cancer genomics consortia, the 1000genomes[4] and the One Million Genomes projects[5], the $10 million genome prize[6] and the remarkable drop in the cost of genome sequencing[7] will soon mean that the current paradigm in which data researchers download all the data, extract the interesting pieces and remove the rest, will no longer be feasible [7, 1]. Advances in statistical methods for analysing cancer genomics [13, 6] further emphasizes the need to enable smooth online data collection and aggregation. As pointed out in [2] "Large-scale genome characterization efforts involve the generation and interpretation of data at an unprecedented scale, which has brought into sharp focus the need for improved information technology infrastructure and new computational tools to render the data suitable for meaningful analyses."

TCGA data has been widely used in the literature (over 350 publications[8]), but mostly in its raw form and without integration beyond a single type of molecular information [10, 12, 8, 4]. Deus et al. [3] developed an infrastructure using Simple Sloppy Semantic Database (S3DB) management model to expose clinical, demographic and molecular data elements generated by TCGA as a SPARQL endpoint. More recently, Robbins et. al [11] developed an engine to continuously index and annotate the TCGA files using JavaScript in conjunction with RDF, and the SPARQL query language. However, both [3] and Robbins et. al [11] provide only file level provenance annotations without providing structured access to actual contents of the files.

A scalable and robust solution is therefore a critical requirement, whereby researchers can obtain the slice of the big data they are interested in by submitting a structured query to a federated service. In addition to the very large semi-structured experimental results datasets available through TCGA and related projects, there is a significant amount of unstructured and structured biomedical data available on the web, which is critical towards annotating and integrating those experimental results. Remote query processing and virtual data integration, i.e. transparent on-the-fly-view creation for the end user, can provide a scalable solution to both challenges. Currently, due to the majority of data being available in text form, it is impossible to query the contents of a particular file or to enable virtual data integration from TCGA data sources. Indeed, the growth of TCGA initiative should also be considered for a scalable solution[9]. We addressed this problem by applying Semantic Web technologies to semi-structured level 3 TCGA data. We converted this data into resource description framework (RDF) data and linked it to the Linked Open Data Cloud so as to make it easy to query. The data can be accessed freely via SPARQL endpoints.

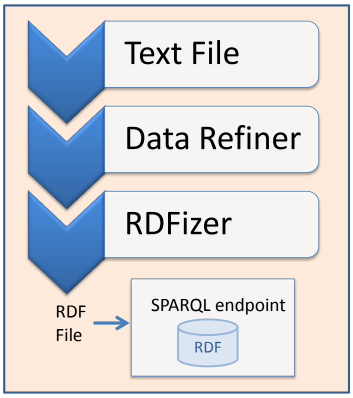


Figure 1: TCGA text to RDF conversion process

2. TCGA RDFIZATION AND LINKING

In this section, we explain the text to RDF conversion process and the linking of the resulting RDF files to the LOD Cloud.

TCGA RDFization

The TCGA text to RDF conversion process is shown in Figure 1. Given a TCGA text file, the Data Refiner selects the specific fields[10] necessary for traditional molecular analysis algorithms. This step is necessary to restrict the size of the resulting RDF according to what we expect will be the most useful results. Finally, the refined text file is send to the RDFizer which generates the resulting RDF file in N3 format so that it can be loaded into any triple store, such as Virtuoso or Sesame.

As an example of the efficient space consumption feature of our RDFization, it is worth noting that original text size (20.63 GB) from the TCGA lung tumour (LUSC) is reduced to 5.75 GB after passing through the Data Refiner and the final RDF files, after passing through RDFizer, only take 20.5 GB to represent 927 million triples. After uploading these files to a virtuoso SPARQL endpoint, the total space consumption is 54 GB. The increase in size (approx. double) is caused by the different indexes created by the virtuoso server for fast data retrieval.

The statistics of the RDFization of the top 10 tumours with the smallest data files is given in Table 1. Given that we have produced a total of 7.34 billion triples for these tumours, we can estimate that entire TCGA level 3 data will result in over 30 billion triples. Our Linked representation of TCGA (to the best of our knowledge) thus promises to be the largest dataset available on the LOD cloud[11].

[4] http://www.1000genomes.org/

[5] http://www.genomics.cn/en/navigation/show_navigation?nid=5658

[6] http://in.reuters.com/article/2012/07/24/us-science-genome-prize-idINBRE86M02G20120724

[7] http://www.genome.gov/sequencingcosts/

[8] TCGAPublications:http://cancergenome.nih.gov/researchhighlights/leadershipupdate/ZhangTCGAStats

[9] http://tcga.github.io/Roadmap

[10] https://code.google.com/p/topfed/wiki/SelectedFields

[11] http://lod-cloud.net/state/

Tumor Type	Original Size(GB)	Refined Size (GB)	RDFized Size (GB)	Triples (Million)
Cervical (CESC)	8.75	2.44	8.86	400.19
Rectal adenocarcinoma (READ)	8.07	2.25	9.04	413.31
Papillary Kidney (KIRP)	10.40	2.90	10.4	469.65
Bladder cancer (BLCA)	12.16	3.39	12.3	556.38
Acute Myeloid Leukemia (LAML)	14.85	4.14	15.1	684.05
Lower Grade Glioma (LGG)	17.08	4.76	17.1	778.82
Prostate adenocarcinoma (PRAD)	18.05	5.03	18.1	821.01
Lung squamous carcinoma (LUSC)	20.63	5.75	20.5	927.08
Cutaneous melanoma (SKCM)	23.22	6.47	23.2	1050.94
Head and neck squamous cell(HNSC)	27.6	7.69	27.5	1245.37

Table 1: Top 10 small (size) TCGA tumours statistics

Result	Target	Class	# links
Methylation	HGNC	Chromosomes	97,530
Methylation	OMIM	Chromosomes	14,407,269
Gene expression	HGNC	Chromosomes	86,052
Gene expression	OMIM	Chromosomes	12,535,829

Table 2: Links for the methylation of a single patient

Linking TCGA to the Linked Open Data Cloud

The fourth design principle behind Linked Data is the provision of links to other data sources. By these means, central tasks such as cross-ontology question answering, data integration and data analytics can be facilitated. Yet, the sheer size of bio-medical knowledge available on the Linked Data Cloud and of TCGA knowledge base itself makes it impossible to use manual linking to provide such cross-knowledge-base links from TCGA to other data sources. We made use of the LIMES framework[12] to compute links between TCGA and knowledge bases. LIMES [9] is a framework for link discovery that provides time-efficient implementations of several string and numeric similarity and distance measures. All the TCGA experimental results are reported with regards to a gene or a chromosome. Given that genes and chromosomes have dedicated IDs that are used across several knowledge bases; we used LIMES exactMatch measure for linking. As such, we focused this work on linking patient data from TCGA (and its reported genetic results) with knowledge bases which describe genes and chromosomes. In particular, we linked TCGA to HGNC[13], OMIM[14] and Homologene[15]. Tables 2 and 3 provides an excerpt of the links generated for the TCGA dataset while Listing 1 provides an excerpt of the specifications used for linking.

TCGA Data Workflow

TCGA data can be organized as a three-layer architecture in which layer 1 contains patient data, layer 2 consists of clinical information and layer 3 contains results for different samples of a patient. Each type of data was assigned to a different class in the RDFized version as depicted in the diagram in Figure 2. In the next section we will describe some use cases where this data is applicable and illustrate the advantages of its RDFization as compared to raw experimental result text files.

Source	Target	Class	# links
DNA27	HGNC	Genes	23,181
DNA27	Homologene	Genes	27,654
DNA27	OMIM	Genes	15,171
DNA450	Homologene	Genes	489,643
DNA450	OMIM	Genes	212,284
DNA27	HGNC	Chromosomes	108,662
DNA27	OMIM	Chromosomes	16,039,535

Table 3: Links for the lookup files of TCGA

Listing 1: Excerpt of the LIMES link specification for linking TCGA and Homologene

```
<SOURCE>
 <ID>TCGA</ID>
 <ENDPOINT>dna_methylation450_Lookup.nt</
     ENDPOINT>
 <VAR>?x</VAR>
 <PAGESIZE>-1</PAGESIZE>
 <RESTRICTION>?x rdf:type tcga-
     schema:dna_methylation450_lookup</
     RESTRICTION>
 <PROPERTY>tcga-schema:Gene_Symbol AS
     lowercase</PROPERTY>
 <TYPE>N-TRIPLE</TYPE>
</SOURCE>
<TARGET>
 <ID>homologene</ID>
 <ENDPOINT>http://homologene.bio2rdf.org/
     sparql</ENDPOINT>
 <VAR>?y</VAR>
 <PAGESIZE>10000</PAGESIZE>
 <RESTRICTION>?y a
     homologene:HomoloGene_Group</
     RESTRICTION>
 <PROPERTY>homologene:has_gene_symbol AS
     lowercase</PROPERTY>
</TARGET>
<METRIC>exactmatch(x.tcga-
    schema:Gene_Symbol,
   y.homologene:has_gene_symbol)</METRIC>
<ACCEPTANCE>
 <THRESHOLD>1</THRESHOLD>
 <FILE>dna_450_homologene_accepted.nt</
     FILE>
 <RELATION>tcga-schema:Homologene</
     RELATION>
</ACCEPTANCE>
```

3. CANCER TREATMENT- USE CASES

In Figure 3, we outline our final goal of linked TCGA Atlas i.e. on-the-fly data collection, analyse it and use rele-

[12] http://limes.sf.net

[13] http://hgnc.bio2rdf.org/sparql

[14] http://omim.bio2rdf.org/sparql

[15] http://homologene.bio2rdf.org/sparql

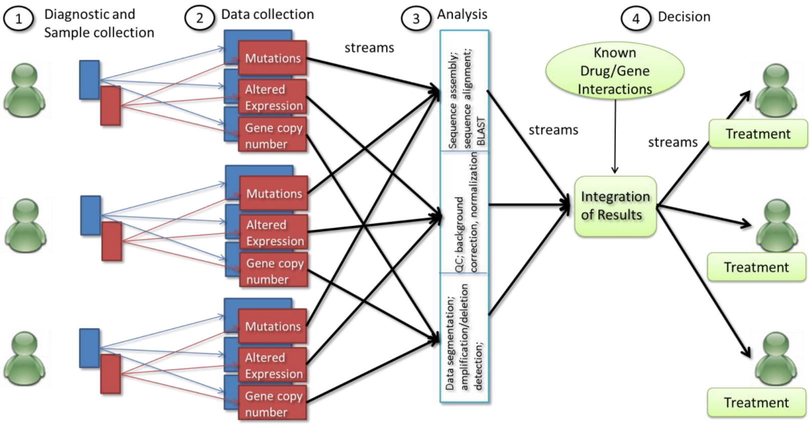


Figure 3: An overview of the pipeline for personalised cancer treatment

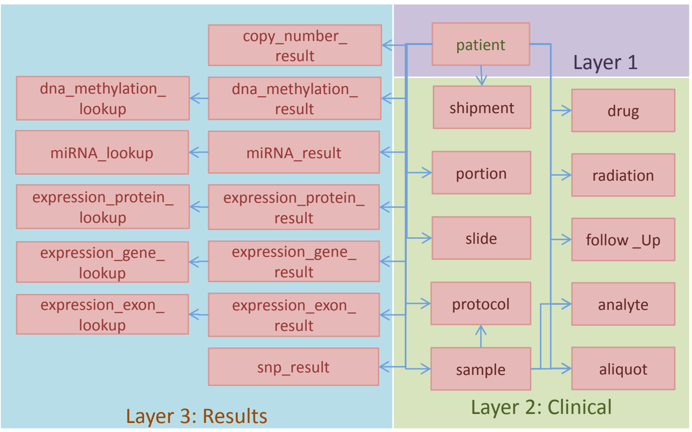


Figure 2: TCGA class diagram of RDFized results

vant data for patient treatment. Stakeholders involved in this process include patients (the primary data providers), physicians, bioinformatics experts and statisticians. Several steps are included in this process:

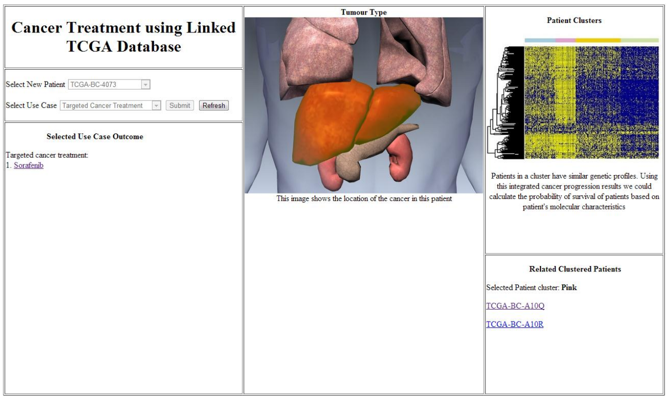
1. Diagnostic and sample collection: Cancer patients around the world are asked for consent towards donating samples for the project.
2. Data collection: Each of the samples is analysed using several molecular techniques for detecting common molecular events in cancer.
3. Analysis: Each type of data needs to be analysed separately according to the platform used. For some of the analysis (e.g. gene expression), batches of patients must be analysed together to enable normalization.
4. Decision: Once the results are analysed for each batch, they are integrated with other data such as known gene/drug interaction.

Currently this process is performed manually and is thus inefficient and error-prone, forcing bioinformatics experts to regularly check for new patient data or new files with data, download the data from multiple endpoints, rerun the analysis and submit the result somewhere where the physician can access and associate it with the patient diagnostic information. Due to such manual process, two critical issues occur: 1) the linking between the patient and his/her genome is lost because the clinical information cannot be made public and 2) statisticians are not encouraged to maintain provenance of models and parameters used to analyse the data, leading often to serious, and expensive mistakes. Hereafter, we describe use cases where automated pipelines will help researchers in improved and backtracked for reproducibility of results. Further, the linking of TCGA with LOD datasets will enable us to further explore the use case outcomes (e.g drugs) in the existing LOD datasets such as HGNC, OMIM, Drugbank and NCBI. The demo of the use cases discussed below is available at `http://linkeddatacup-demo.deri.ie/`.

3.1 Targeted Cancer Treatment

The main question addressed in this use case is whether a specific drug can be used to treat a tumour given the genomic data of those tumour patients. An example for this use case can be seen in Breast Cancer, patients having mutations in BRCA1 and BRCA2 genes are highly susceptible to Breast Cancer and have varying treatment compared to patients without mutations in these genes. Another example in Breast Cancer is HER2-positive breast cancer, which warrants different treatment from HER2-negative breast cancers. Many such studies have been done to find clinical subtypes of different cancers for targeted treatment.

Given that these genetic mutations only occur in a handful of cases, in order for these kinds of studies to have strong statistical predictability, there is need for a very large sample size of cancer patients data - much more than what a single hospital can produce. The TCGA project aimed at assembling this very large cancer cohort but, to the best of our knowledge, the lack of a structured representation of the data proposed in the report prevented these large correlations to be derived. Given the integration of respective cancer omics data one can use different bioinformatics methods to find relevant genomic profiles in particular tumour type



Cancer Treatment using Linked TCGA Database

Select New Patient TCGA-BC-4073

Select Use Case Targeted Cancer Treatment

Selected Use Case Outcome

Targeted cancer treatment:
1. Sorafenib

Tumour Type

This image shows the location of the cancer in this patient

Patient Clusters

Patients in a cluster have similar genetic profiles. Using this integrated cancer progression results we could calculate the probability of survival of patients based on patient's molecular characteristics

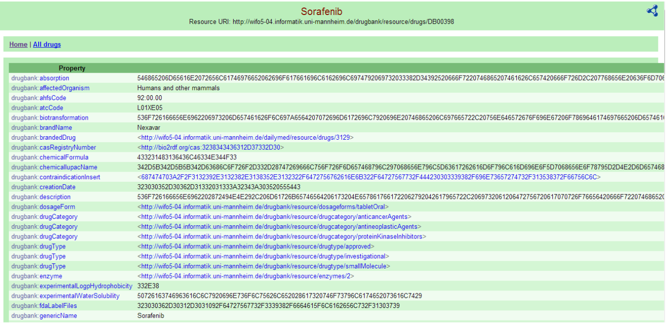
Related Clustered Patients

Selected Patient cluster: **Pink**

TCGA-BC-A10Q

TCGA-BC-A10R

Figure 4: Screenshot of the Targeted Cancer Treatment



Sorafenib
Resource URI: http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/drugs/DB00398

Home | All drugs

Property	
drugbank:absorption	546865206D65616E2072656C61746976652062696F617661696C6162696C6974792069732033382D34392520666F7220746865207461626C657420666F726D2C207768656E20636F6D706
drugbank:affectedOrganism	Humans and other mammals
drugbank:ahfsCode	92:00.00
drugbank:atcCode	L01XE05
drugbank:biotransformation	536F726166656E6962206973206D657461626F6C697A6564207072696D6172696C7920696E20746865206C697665722C20756E646572676F696E67206F7869646174697665206D657461
drugbank:brandName	Nexavar
drugbank:brandedDrug	<http://wifo5-04.informatik.uni-mannheim.de/dailymed/resource/drugs/3129>
drugbank:casRegistryNumber	<http://bio2rdf.org/cas:3238343436312D37332D30>
drugbank:chemicalFormula	433231483136436C46334E344F33
drugbank:chemicalIupacName	342D5B342D5B5B342D63686C6F726F2D332D28747269666C756F726F6D657468796C297068656E796C5D63617262616D6F796C616D696E6F5D7068656E6F78795D2D4E2D6D657468
drugbank:contraindicationInsert	<687474703A2F2F3132392E3132382E3138352E3132322F6472756762616E6B322F64727567732F444230303339382F696E73657274732F313538372F66756C6C>
drugbank:creationDate	323030352D30362D31332031333A32343A303520555443
drugbank:description	536F726166656E6962202872494E4E292C206D61726B65746564206173204E6578617661722062792042617965722C2069732061206472756720617070726F76656420666F7220746865
drugbank:dosageForm	<http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/dosageforms/tabletOral>
drugbank:drugCategory	<http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/drugcategory/anticancerAgents>
drugbank:drugCategory	<http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/drugcategory/antineoplasticAgents>
drugbank:drugCategory	<http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/drugcategory/proteinKinaseInhibitors>
drugbank:drugType	<http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/drugtype/approved>
drugbank:drugType	<http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/drugtype/investigational>
drugbank:drugType	<http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/drugtype/smallMolecule>
drugbank:enzyme	<http://wifo5-04.informatik.uni-mannheim.de/drugbank/resource/enzymes/2>
drugbank:experimentalLogpHydrophobicity	332E38
drugbank:experimentalWaterSolubility	50726163746963616C6C7920696E736F6C75626C652028617320746F73796C6174652073616C7429
drugbank:fdaLabelFiles	323030362D30312D3031092F64727567732F3339382F6664615F6C6162656C732F31303739
drugbank:genericName	Sorafenib

Figure 5: Screenshot of the effective drug for Targeted Cancer Treatment taken from DrugBank SPARQL endpoint `http://wifo5-04.informatik.uni-mannheim.de/drugbank/page/drugs/DB00398`

having specific drug effect as clinical variable of interest.

As an example, we present in Listing 2 a query that retrieves all patients having breast tumour, together with information about their treatment and relapse. The results of this query will be further analysed using statistical tools to find alternation in HER2 and ER genes in patients. It can be seen in Figure 4 that a patient is selected for targeted cancer treatment where list of drugs that are specific for such type of cancer are displayed as results. The output of this use case is based on strong correlation between the genomic data of the selected input patient and previously collected patients of same cancer. The patient clusters show the number of clinical subtypes of cancer that can be found in the collected patients genomic data. It can also be seen to which cluster (pink in this example) does the input patient belong. Based on this information specific drug (i.e Sorafenib) is suggested for that patient's treatment. The information about the selected drug can further be explored by using the LOD data sets such as DrugBank as shown in Figure 4.

3.2 Mechanism-based Treatment

The main question addressed in this use case is whether a combination of drugs can be applied to treat a specific tumour effectively. Cancer can be regarded as a series of aberrant genetic events leading to an uncontrollable cell growth.

Listing 2: Use case 1,2 SPARQL query

```
 Select ?patient ?mean
where
{
?uri tcga:tumour_type "BRCA".
?uri tcga:bcr_patient_barcode ?patient.
?patient rdf:type  tcga:expression_gene_results.
?patient tcga:gene_symbol "HER2","ER".
?patient tcga:scaled_estimate ?mean
}
```

Listing 3: Querying LOD DrugBank

```
SELECT ?drugname
WHERE
{
?patient rdf:type tcga:expression_gene_results.
?patient tcga:gene_symbol ?targetname .
?patient tcga:scaled_estimate ?mean.
FILTER (?mean > Threshold)
?drug drugbank:target ?target.
?drug drugbank:genericName ?drugname .
?target drugbank:synonym ?targetname .
FILTER REGEX (?targetname, "HER2|estrogenreceptor|ERBB2",
      "i")
}
```

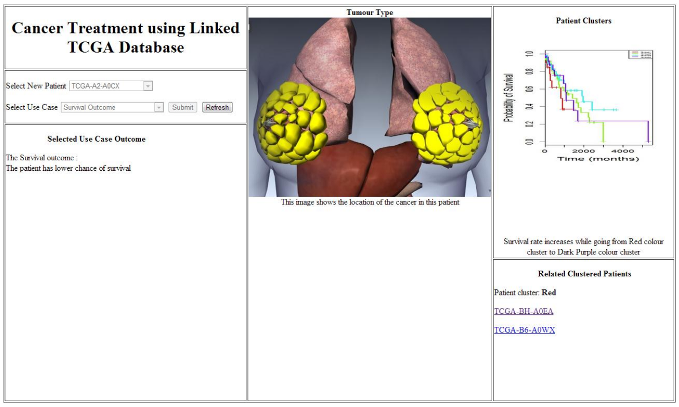
As mentioned in section 3.1, there are drugs specific for certain genetic events, which can be prescribed to patients differentially, making their treatment personalized. Furthermore, it is often the case that patients on one drug often relapse (because the cancer has become resistant to that drug). Thus, a combination of drugs - either prescribed together or in sequence - might be necessary for effective treatment of cancer. The patient list from statistical analysis of use case 1 results can be used to detect which patients are sensitive to the drug Transtuzumab (which targets the HER2 gene) and intercept with patients that are sensitive to the drug Tamoxifen (which targets the gene estrogen receptor) using the information from other LOD sources targeting drugs such as DrugBank as shown in Listing 3. The threshold in Listing 3 is obtained from the statistical analysis of use case 1 results. To explore such areas, integration of cancer omics data such as provided here was of great importance to produce statistically significant results.

3.3 Survival Outcome

The main question addressed in this use case is whether a mathematical model can be built on the patients tumour omics and clinical data in order to detect signs of tumour given the genomic profile of a future patient. It is well known that the treatment of early stage tumours has a much higher success rate than that of late-stage tumours. The classification of tumour patients based on the genetic biomarkers along with patients cancer omics and clinical data can be a powerful predictive tool with increasing tumour patients

Listing 4: Use case 3 SPARQL query

```
 Select ?patient ?mean
where
{
 ?uri tcga:tumour_type "BRCA".
 ?uri tcga:bcr_patient_barcode ?patient.
 ?patient rdf:type  tcga:clinical.
 ?patient tcga:tumour_stage ?tumour_stage.
 ?patient tcga:age_at_initial_patalogical_diagnosis ?age.
 ?patient tcga:relevant_biomarker "BRCA1","CDKN2A","CDH1".
 ?patient tcga:beta_value ?mean
}
```



Cancer Treatment using Linked TCGA Database

Select New Patient: TCGA-A2-A0CX

Select Use Case: Survival Outcome

Selected Use Case Outcome

The Survival outcome :
The patient has lower chance of survival

Tumour Type

This image shows the location of the cancer in this patient

Patient Clusters

Survival rate increases while going from Red colour cluster to Dark Purple colour cluster

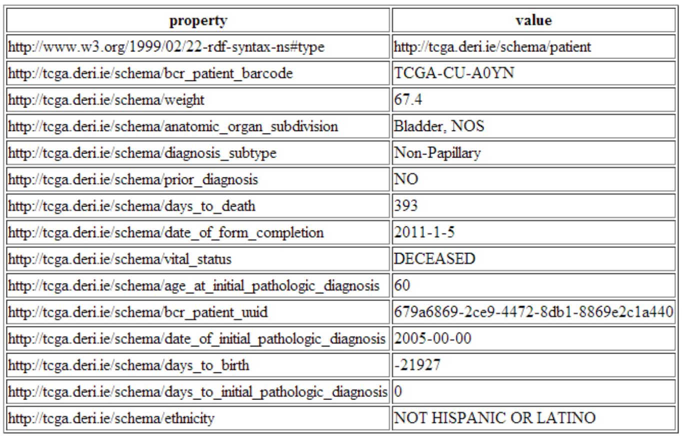
Related Clustered Patients

Patient cluster: **Red**

TCGA-BH-A0EA

TCGA-B6-A0WX

Figure 6: Screenshot of the Survival Outcome



property	value
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://tcga.deri.ie/schema/patient
http://tcga.deri.ie/schema/bcr_patient_barcode	TCGA-CU-A0YN
http://tcga.deri.ie/schema/weight	67.4
http://tcga.deri.ie/schema/anatomic_organ_subdivision	Bladder, NOS
http://tcga.deri.ie/schema/diagnosis_subtype	Non-Papillary
http://tcga.deri.ie/schema/prior_diagnosis	NO
http://tcga.deri.ie/schema/days_to_death	393
http://tcga.deri.ie/schema/date_of_form_completion	2011-1-5
http://tcga.deri.ie/schema/vital_status	DECEASED
http://tcga.deri.ie/schema/age_at_initial_pathologic_diagnosis	60
http://tcga.deri.ie/schema/bcr_patient_uuid	679a6869-2ce9-4472-8db1-8869e2c1a440
http://tcga.deri.ie/schema/date_of_initial_pathologic_diagnosis	2005-00-00
http://tcga.deri.ie/schema/days_to_birth	-21927
http://tcga.deri.ie/schema/days_to_initial_pathologic_diagnosis	0
http://tcga.deri.ie/schema/ethnicity	NOT HISPANIC OR LATINO

Figure 7: Screenshot of the clinical TCGA related patient breast cancer data

sample size. For this use case, the query given in Listing 4 selects patients with available tumour stage along with relevant clinical variables and selects methylation biomarkers which are correlated to the tumour stage.

The results from the Listing 4 are further sent for analysis tools to classify the patients in to relevant clinical clusters based on the tumour stage. The need for integrating the tumour patients' data for this use case is fulfilled by this project. It can be seen in Figure 6 that patients are divided in to clusters based on statistical modelling of gene expression data to patient survival time. Different clusters indicate for different survival times and patients in a certain cluster have more or less similar tumour stage. The output of this use case will be predicting the survival rates for the given input patient based on the patient's gene expression data. The chance of survival depends upon the cluster the input patient belongs; with red cluster has the lowest and purple has the highest chance of survival. In the demo screen shot, the input patient belongs to red cluster indicating that it has lowest chance of survival. Furthermore, patients who belong to the same category of lower survival rates are shown and the clinical data of these patients can be seen in Figure 7 which has details of patients cancer type, drugs used, date of admission and so on.

4. ACKNOWLEDGMENTS

This research has been supported by the European Commission under Grant No. FP7-Granatum: RE7098, FP7-GeoKnow Grant Agreement No. 318159 and by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-II) and Grant No. SFI/12/RC/2289 (INSIGHT). It is further supported by Foundation Medicine Inc. Cambridge, MA.

5. REFERENCES

[1] G. Bell, T. Hey, and A. Szalay. Beyond the data deluge. pages 1297–1298, 2009.

[2] L. Chin, W. C. Hahn, G. Getz, and M. Meyerson. Making sense of cancer genomic data. *Genes & development*, 25(6):534–555, 2011.

[3] H. F. Deus, D. F. Veiga, P. R. Freire, J. N. Weinstein, G. B. Mills, and J. S. Almeida. Exposing the cancer genome atlas as a sparql endpoint. *Journal of Biomedical Informatics*, 43(6):998 – 1008, 2010.

[4] F.-H. Hsu, E. Serpedin, T.-H. Hsiao, A. J. Bishop, E. R. Dougherty, and Y. Chen. Reducing confounding and suppression effects in tcga data: an integrated analysis of chemotherapy response in ovarian cancer. *BMC Genomics*, 13(Suppl 6):S13, 2012.

[5] T. J. Hudson, W. Anderson, A. Aretz, A. D. Barker, C. Bell, R. R. Bernabé, M. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

[6] J. Jeong, L. Li, Y. Liu, K. Nephew, T. Huang, and C. Shen. An empirical bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer. *BMC medical genomics*, 3(1):55, 2010.

[7] J. Karlsson, O. Torreño, D. Ramet, G. Klambauer, M. Cano, and O. Trelles. Enabling large-scale bioinformatics data analysis with cloud computing. In *ISPA*, pages 640–645, 2012.

[8] H. S. Kim, J. D. Minna, and M. A. White. Gwas meets tcga to illuminate mechanisms of cancer predisposition. *Cell*, 152(3):387–389, 2013.

[9] A. Ngomo. On link discovery using a hybrid approach. *J. Data Semantics*, 1(4):203–217, 2012.

[10] H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat, R. G. Verhaak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. V. D. Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, and K. Aldape. Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, 17(5):510 – 522, 2010.

[11] D. E. Robbins, A. Grueneberg, M. M. Tanik, H. F. Deus, and J. S. Almeida. A self-updated roadmap of the cancer genome atlas. *Bioinformatics*, 2012.

[12] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

[13] K. D. Siegmund. Statistical approaches for the analysis of dna methylation microarray data. *Human genetics*, 129(6):585–595, 2011.