Linked Functional Annotation For Differentially Expressed Gene (DEG) Demonstrated using Illumina Body Map 2.0

Alokkumar Jha , Yasar Khan, Muntazir Mehdi, Aftab Iqbal, Achille Zappa, Ratnesh Sahay, and Dietrich Rebholz-Schuhmann

Insight Centre for Data Analytics, NUI Galway, Ireland
Galway
{alokkumar.jha,yasar.khan,aftab.iqbal,achille.zappa,muntazir.mehdi,
ratnesh.sahay,rebholz}@insight-centre.org

Abstract. Semantic Web technologies are core for the integration of disparate data resources. It can be used to exploit data from next generation sequencing (NGS) for therapeutic decisions regarding cancer. In this manuscript, we describe how different data resources, which inform on the expression of specific genes in a tissue and its variants, can be brought together to indicate a risk for tissue-specific cancer for NGS data. This approach can be used to judge patient genomic data against public reference data resources.
The TCGA and COSMIC repositories are being processed to connect and query information concerning the expression of genes, copy number variants (CNV), and somatic mutations. We annotated sets of differential expression data provided from the Illumina Body map 2.0 (HBM) concerning 16 different tissue types and identify genes with an RPKM (Reads Per Kilobase of transcript per Million mapped reads) value greater than 0.5 as measure indicating an associated risk for cancer. Thus, the differential expressed genes from HBM can be associated with a tissue type and gene expressions in COSMIC and TCGA leading to a potential biomarker for that particular tissue specific cancer. In the case of ovarian cancer, we retrieved the genomic positions (loci) and the associated genes of potential biomarker candidates, and suggest that this approach and platform can serve future studies well.
Altogether, the presented linked annotation platform is the first approach to represent the COSMIC data in an RDF format and to link the data with the TCGA datasets. The proposed approach enriches mutations by filling in missing links from COSMIC and TCGA datasets which in turn helped to map mutations with associated phenotypes.

Keywords: Differentially Expressed genes(DEG),Linked data, Clinical genomics,Copy Number Variation (CNV)

1 Introduction

Next Generation Sequencing (NGS) technologies open new diagnostic and therapeutic ways for cancer research. However, the resulting high-throughput sequencing data has to be processed in complex data analytics pipelines including

annotation services. Unfortunately, there is not yet a well-integrated platform available for both clinical and translational [12] research to fulfill these annotation and analytical tasks. In addition, the large volumes of NGS data poses another challenge, since the computational infrastructure for the biological interpretation will have to cope with very large quantities of data originating from clinical facilities. Last, but not least, the functional annotation of genomics data for cancer has to take tissue specificity into consideration and thus has to avoid ambiguity while aggregating clinical outcomes from disparate resources. In this paper we focus on exploring gene expression patterns across different cancer and tissue types. Our experiments are based on semantic integration of gene expression, CNV, complete mutation data from two disjoint resources, i.e. COSMIC[1] and TCGA[2]. By doing this we can assist in variant and mutation prioritization using 16 different tissue types given by the *Illumina Body Map 2.0* and evaluated in a case study for *Ovarian cancer*.

In order to link and retrieve patterns of a gene and tissue specific information from various cancer mutation (TCGA) and database with global mutation list and mutation type (COSMIC), we encountered the following three challenges: (i) to transform heterogeneous data repositories and their storage formats into standard RDF; (ii) to discover associations (aka. links) by finding specific patterns (i.e. correlations) for a gene with regards to CNV, mutation and its gene expression data sets; and (iii) to query in a scalable way the large volume and frequently updating datasets covering 16 different tissue types and the gene expression data from different repositories.

The experiments conducted in this paper is aligned to the transcriptome study based on the Human Body Map 2.0 (HBM)[3] from Illumina which covers the following tissues: adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells. The HBM provides gene specific information across one or more tissue types and intends to support the identification of potential biomarker for targeted therapy. In this study our results not only depicts novel biological outcomes but also provides a linked annotation framework that assimilates clinical outcomes from related data repositories.

The rest of the paper is structured as follows: Section 2 motivates our working scenario exploring on the HBM use case and the annotation databases; Section 3 presents the methodology and architecture of the proposed functional annotation framework; Section 4 gives an evaluation of the functional annotation framework; Section 6 presents the related work in linking the TCGA repository and Section 7 draws the conclusion from our work.

2 Motivation

In order to understand the outbreak of disease, in particular cancer, it is one approach to compare normal and diseased tissue samples to interpret the changes in the expression patterns of the genes with regards to the observed disease status.

[1] `http://cancer.sanger.ac.uk/cosmic`

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

[3] `https://www.ebi.ac.uk/gxa/experiments/E-MTAB-513`

In our case, HBM serves the purpose to identify similarities in gene expression patterns using the studies across different tissue types, where HBM discloses the similarities between human tissues on the molecular and genetic level. Due to overlaps between cancer behaviours, progression and mutated genes, we have annotated top 100 genes distilled by our filtering criteria with COSMIC to explore previously observed studies from TCGA database, e.g., somatic mutation, genomic loci and other mutations linked to these genes retrieved from healthy tissues.

Human Body Map (HBM) 2.0 from Illumina: HBM covers data from transcriptome studies for 16 tissue types (see above). Samples for these 16 tissue types have been processed, aligned and finally expression level have been determined [1]. Sequencing has been performed to provide both paired-end and single-end libraries (read-length of 50bp and 75bp). Therefore, the data processing platform requires a list of differentially expressed genes as input, which is the outcome of the RNA seq data analysis pipeline.

The gene expression data extracted from HBM samples returns a very large list of more than 52000 genes. For data processing reasons we chose to reduce the list and therefore defined the cut off for each *RPKM* value according to the method suggested by Sandberg et. al[8]. As a result, the data for each tissue type includes both the coverages and the *RPKM* values as the corresponding expression level. In addition, the RNA seq data set provides further relevant data such as CNV, fusion genes, structural variation, differentially expressed genes, novel mutations, splice junctions and trascriptome variations [4]. Identifying the associations and relations between these datasets, i.e. the logical connections, enables further insightful research into the cancer disease the biological and clinical interpretation of given data.

Annotation databases (COSMIC & TCGA): The main focus of this work is the identification of patterns for cancer mutations (given by TCGA) and globally known mutations and their types (given by COSMIC) for selected differentially expressed genes across different tissue types. Figure 1 shows the correspondences, i.e., the associations or connections that have been established between the TCGA and COSMIC databases for this purpose. For this task our primary concern have been the associations between the CNV, the known mutations and the gene expression data.

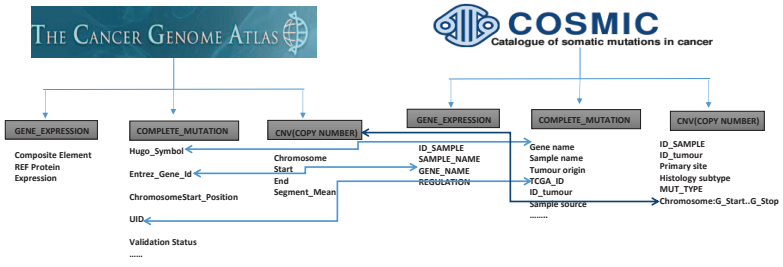

Fig. 1: Links between COSMIC and TCGA repositories

As part of this work, specific basic curation for data refinements have been performed: we had to identify instances to link two databases or a couple of

events within the databases (see fig. 1). For example, MUTATION and GENE_EXPRESSION data in COSMIC could be linked to GENE_NAME but CNV had SAMPLE_IDs as expected. Later we used SAMPLE_IDs after first iteration with GENE_NAME. Also, chr:start_end position and GENE_NAME were used to link COSMIC and TCGA (see green arrows in Figure 1). The RDFized version (see section 3.1) has kept this redundancy problem to have *FDR* rate as low as possible.

3 Methodology & Architecture

The annotation architecture is summarized in Figure 2 showing all three major components. First, the RDFization component that generates Linked Data from the TCGA and COSMIC databases leading to several SPARQL endpoints for public use. Second, the linking component that searches and discovers correspondences between selected datasets (TCGA_variants, COSMIC_diff_expression, COSMIC_Mutation, etc.). The links discovered by this component have an effect on the efficiency in the source selection, on the query planning, and on the overall query execution in a decentralized setting. Third, the scalable query federation component: it a single-point-of-access through which distributed data sources can be queried in concerto.

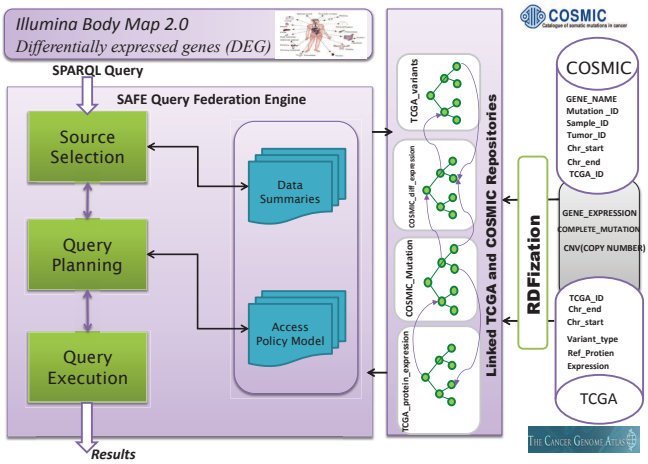

Fig. 2: Linking and functional annotation of gene expression data

The scalable query federation is based on the SPARQL query federation engine called SAFE [7], which has been developed for accessing distributed clinical trial repositories. SAFE has been adapted to improve the efficient integration of data from the different TCGA and COSMIC SPARQL endpoints. More specifically, SAFE makes use of a favourable distribution of data to reduce the number of sources required for processing federated SPARQL queries (without compromising recall). This approach is based on the principle that integrated data sources allow querying of multiple data sources in a single search, independently of their status being distributed or centralized, whereas traditional methods of data integration rather map the data models to a single, unified, model. Such

methods tend to resolve syntactic differences between models, but do not address possible inconsistencies in the concepts defined in those models. Semantic integration resolves the syntactic heterogeneity present in multiple data models as well as the semantic heterogeneity among similar concepts across those data models.

3.1 RDFization

COSMIC raw data files are delivered as tab separated text (tsv) and are being processed with the COSMIC RDFizer tool that generates the N3 triples for the SPARQL endpoint and statistical information related to the data. Only three types of data have been included, i.e., gene expression data, gene mutation and CNV data. Table 1 shows the overall statistics of the RDFization: row 1 represents for the COSMIC gene expression data the number of records (column 2), it's size (column 4), the corresponding triples generated (column 2) and again it's size. The other two rows represent the same type of data for the COSMIC gene mutation and CNV data. A total of 154 million records has been RDFized, producing approximately 1.2 billion triples. Row 5 represents the statistics for the RDF version of TCGA-OV (TCGA Ovarian), which forms a subset of the linked TCGA[4] data. The RDF file for the COSMIC data can be made available inline with the COSMIC data policy.

Table 1: COSMIC data statistics

No.	Data	Records	Triples	Original Size	RDFized Size
1	COSMIC GE	149 M	1185 M	7.5 GB	20 GB
2	COSMIC GM	3.7 M	84 M	916 MB	1.44 GB
3	COSMIC CNV	0.9 M	11 M	82 MB	161 MB
4	Total	154M	1.28 B	8.5 GB	22 GB
5	TCGA-OV	18M	100 M	2.15 GB	5 GB

3.2 COSMIC and TCGA linking

The main integration is based on `owl:sameAs` constructs as can be seen in the listing 1.1 where two COSMIC sample ids have been identified as being identical to two TCGA patient bar code ids. These links are at the core of facilitating data integration and the data analysis tasks.

Listing 1.1: COSMIC and TCGA Linking Example

```
<Link-1>
  <Source>COSMIC</Source>
  <Target>TCGA-OV</Target>
  <link>
    http://cosmic.sels.insight.org/schema/ID_Sample/TCGA-13-0920
    owl:sameAs
    http://tcga.deri.ie/TCGA-13-0920
  </link>
  </Link-1>
<Link-2>
  <Source>COSMIC</Source>
  <Target>TCGA-OV</Target>
```

[4] `http://tcga.deri.ie/`

```
  <link>
    http://cosmic.sels.insight.org/schema/ID_Sample/TCGA-24-1850
    owl:sameAs
    http://tcga.deri.ie/TCGA-24-1850
    </link>
</Link-2>
```

3.3 Scalable query federation

SAFE has been developed for accessing sensitive clinical data in data cubes at different locations [7]. Two main changes have been introduced to SAFE for efficiently querying the TCGA and COSMIC SPARQL endpoints. First, standardize RDF query representation: in the initial versions, SAFE issues queries for statistical clinical information stored within distinct names graphs for RDF data cubes [3]. Therefore, the internal query processing (i.e., source selection, query planning, query execution) had to be adapted to query the regular RDFized versions of the TCGA and COSMIC repositories. Second, access control had to be disabled: SAFE imposes restrictions for data-access as a feature (defined as Access Policy Model [7]) while federating queries over multiple clinical site, i.e. imposing the data restrictions for different data repositories. Since, experiments conducted in this paper mainly involve public repositories this feature has been disabled.

The listing 1.2 shows a sample SPARQL query, which federates across COSMIC and TCGA data asking for genomic *loci* of a mutated gene by chromosome start points which then returns the disease metastasis information along with the mutation type. Answering such a query requires the integration of COSMIC with TCGA and merging results from both TCGA and COSMIC, and thus has to make use of query federation. The results for the first four triples in the given query (i.e. `cosmic-s:ID_Sample`, `cosmic-s:Gene_Name`, `cosmic-s:Chrom_start`) are fetched from COSMIC and the results for the next three triples (i.e. `tcga:tcga_id`, `tcga:start`) are fetched from TCGA. To produce the required information, both results are merged on the basis of the last triple which integrates COSMIC with TCGA. Sample results for this query can be seen in Figure 5.

Listing 1.2: Federated SPARQL Query

```
PREFIX cosmic-s: <http://cosmic.sels.insight.org/schema/>
PREFIX tcga: <http://tcga.deri.ie/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT * WHERE {
 ?cosmic_result a cosmic-s:result;
   cosmic-s:ID_Sample ?id_sample; cosmic-s:Gene_Name ?gene;
   cosmic-s:chrom_start ?chrom_s_cosmic.
 ?tcga_result a tcga:result;
   tcga:Id ?tcga_id; tcga:Start ?tcga_chrom_start.
 ?id_sample owl:sameAs ?tcga_id.
}
```

4 Biological questions and annotation results from HBM

We analysed all genes have an RPKM > 0.5 [8] and that are differentially expressed in all tissue types. Figure 3 [2] is a schematic representation which

satisfies all mentioned conditions and delivers 99 genes per query. We have identified potential cancer types based on gene patterns for different tissues and further helped to understand the behaviour of most amplified cancer types.

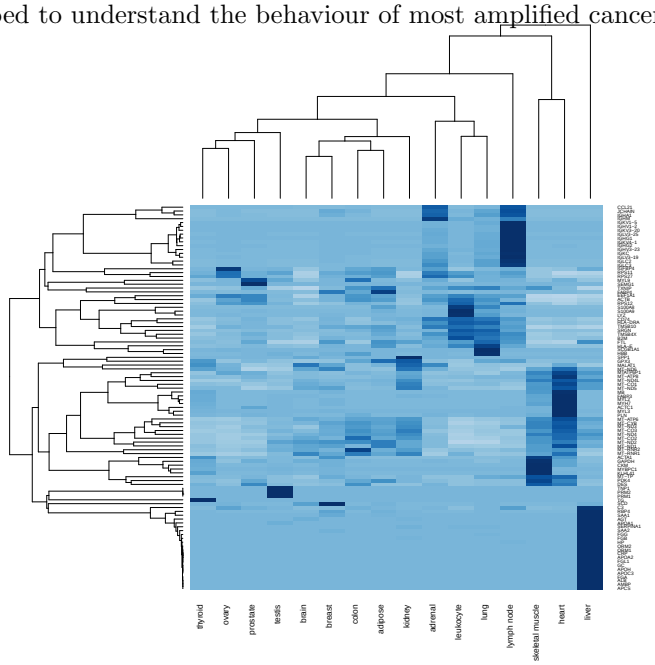

Fig. 3: List of Genes expressed in all tissues and highly expressed.

5 Results

The overall goal of this study is to understand the relevance of mutations and genes along with their associated expression levels measured in data sets for normal tissues e.g. HBM 2.0, and then evaluate (e.g., query) them against the mutations retrieved and linked from somatic and patient specific data e.g COSMIC, TCGA. Further focus of this work is put to the linked annotations where a single query can retrieve all other possibly relevant annotations.

Initially, we have sampled 99 genes that are highly expressed in all 16 tissues as shown in Figure 3 to retrieve their CNV, mutation and gene expression annotations from cBIO portal (for TCGA) and CNV annotator (for COSMIC) to determine current state of the art and provide a baseline comparison for the proposed linked annotation solution. The results for TCGA 4 clearly indicate an elevated distribution of these genes in *uterine* and *ovarian* cancer with a large number of mutations and CNVs. As an outcome, ovarian cancer has been selected as a good candidate for further investigation due to its elevated amplification rate and its multiple repetition in different experiments. Further studies have been retrieved that were conducted to understand the somatic relevance and the *loci(genomic position)* of these genes, and further detailed information for mutations could be retrieved as well. This study demonstrated a focus to genes such as *ACTC1, B2M, CRP, FABP3, FABP4, FGA, FGB,*

GC, MYH7, RPRH2, SLC26A3, TG, TXNIP which form most relevant driver genes transforming healthy human tissues into ovarian cancer. The same 99 genes have been queried against 99 genes from HBM 2.0 to get the results for the somatic mutations in cancer. This repeat annotation will not only provide detailed statistics reported in COSMIC but also a validation for out earlier experiments. Table 2 clearly indicates that locations ***chr1,chr4,chr14*** and genes ***CRP,FGA,FABP3,MYH7*** could be potential genes with high relevance for the development of ovarian cancer.

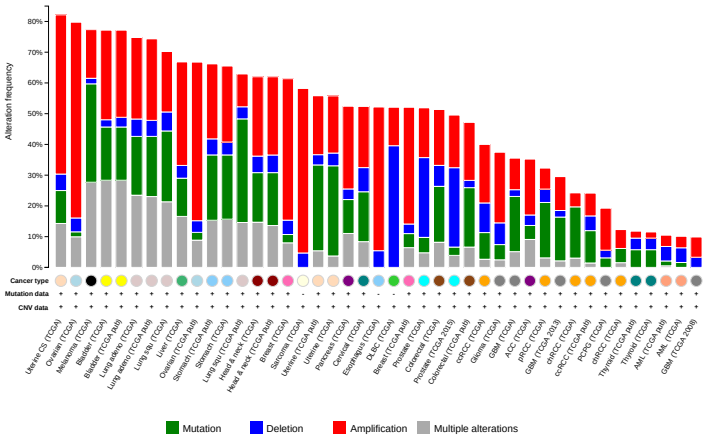

Fig. 4: TCGA query output from cBIO Portal[5]

We now query the same 99 genes from before using the linked annotation mechanism. The result snippet for ovarian cancer is depicted in Figure 5 with detailed functional annotations together with the TCGA ids, again for ovarian cancer.

```
cs:result cs:Sample_Name c:TCGA-13-0920-01;
cs:Gene_Name c:MYH7;
cs:Regulation "over".
cs:chr_no c:1;
cs:chrom_start_m  c:23418303;
cs:chrom_stop_m c:23418303;
cs:chr_no_m c:14;
cs:mut_type "GAIN";
cs:Primary_Site c:ovary;
cs:Primary_Histology c:carcinoma;

ts:result ts:bcr_patient_barcode t:TCGA-13-0920;
ts:beta_value "0.0419..";
ts:chromosome t:14;
ts:chromosome t:1;
ts:start t:1288070;
ts:stop t:1293914;
ts:scaled_estimate "773.555".
```

Fig. 5: Linked annotations for MYH7 - COSMIC

Figure 5 represents the COSMIC and TCGA annotations, respectively. *MYH7* corresponds to chr-1 which is evident from previous annotations and replicated again in our study along with its **TCGA ID:TCGA-13-0920-01**. Its mutation type is primarily the GAIN type of a mutation for chr1 and chr14 which is a dominant mutation with all its regulation of over, under and normally expressed. Translational researchers may want to repeat and re-validate the study for Pubmed ID:1398522 additionally with *beta value* (measure of methylation) of 0.041999536 and scaled estimation (Tumour purity) of 773.555 also supports this gene from the epigenetic point of view. Further multiple genomic locations

will help clinical practitioners to track CNV for the targeted study and which ultimately leads the direction towards a better prognosis.

Table 2: loci information for highly expressed gene in ovarian cancer from HBM 2.0

CHROMOSOME	Mutation Type	Pubmed	GENES
chr1:246407146-246740944	CNV	17122850	CNST, LOC255654, SMYD3, TFB2M
chr1:159683086-159684133	Loss	20364138	CRP
chr1:31842502-31849609	Deletion	20482838	FABP3
chr1:31842502-31849609	Deletion	20482838	FABP3
chr4:155506556-155506859	Insertion	20981092	FGA
chr14:23857082-23886607	Loss	20981092	MIR208A, MYH6, MYH7
chr14:23857092-23886486	Loss	20981092	MIR208A, MYH6, MYH7

6 Related Work

Kandoth et al. [6] performed a cancer study with 12 cancer types to enable logical classifications for the large amoung of data generated by TCGA and ICGC. Saleem et. al. [11] have covered TCGA database with few cancer types and for a limited number of patient data.

Likewise a reduced version of the COSMIC database has been RDFized to explore on the mechanism of TP53 [13] further for CNV explains the linked infrastructure to annotated CNV. The federation platform [10] called "TopFed" is being developed to measure the query execution time on TCGA data set, which then has been further extended to cover the biological outcomes identified from Medline abstracts [9]. Our work covers all CNVs, mutations and gene expression data and has been extended with TCGA for the same type of data, thus forming a proof of concept for an annotation platform that covers comprehensive linked life science data. It is important to note that the work presented in this paper is a preliminary approach for transforming COSMIC into the RDF format and link it with the TCGA datasets.

7 Conclusion

In this paper we have presented a linked data infrastructure for functional annotation which enables querying different types of mutations and genomic alterations to contribute to molecular and clinical insights of cancer by defining most relevant variants and their prioritization. This knowledge could be highly advantageous for a targeted therapy and personalized medicine based on gene expression data. The presented experiments are based on TCGA, COSMIC and HBM 2.0 datasets and have been used to identify sets of genes with relevance for ovarian cancer and with comprehensive set of mutations. Similar studies have to be performed for other cancer types. We have covered CNV, gene expression and mutation data from COSMIC and TCGA (only for ovarian cancer). We have processed 1.2 billion COSMIC triples and 100 million TCGA triples which in turn generated 27 GB of data. In future, this work will be expanded to cover level 1, 2 and 3 along with other datasets from COSMIC to provide in-depth biological insight for each queried gene.

8 Acknowledgment

This publication has emanated from research supported by the research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. Y. W. Asmann, B. M. Necela, K. R. Kalari, A. Hossain, T. R. Baker, J. M. Carr, C. Davis, J. E. Getz, G. Hostetter, X. Li, et al. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer research*, 72(8):1921–1928, 2012.
2. E. bioinformatics institute. Illumina body map 2.0 european bioinformatics institute.
3. J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *WWW*, pages 613–622, 2007.
4. J. J. Crowley, V. Zhabotynsky, W. Sun, S. Huang, I. K. Pakatci, Y. Kim, J. R. Wang, A. P. Morgan, J. D. Calaway, D. L. Aylor, et al. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature genetics*, 2015.
5. J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, 6(269):pl1–pl1, 2013.
6. C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 2013.
7. Y. Khan, M. Saleem, A. Iqbal, M. Mehdi, A. Hogan, A. N. Ngomo, S. Decker, and R. Sahay. SAFE: policy aware SPARQL query federation over RDF data cubes. In *Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences, Berlin, Germany, December 9-11, 2014.*, 2014.
8. D. Ramskold, E. T. Wang, C. B. Burge, and R. Sandberg. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, 5(12):e1000598, 2009.
9. M. Saleem, M. R. Kamdar, A. Iqbal, S. Sampath, H. F. Deus, and A.-C. N. Ngomo. Big linked cancer data: Integrating linked tcga and pubmed. *Web Semantics: Science, Services and Agents on the World Wide Web*, 27:34–41, 2014.
10. M. Saleem, M. R. Kamdar, A. Iqbal, S. Sampath, H. F. Deus, and A.-C. Ngonga. Fostering serendipity through big linked data. *Semantic Web Challenge at ISWC*, 2013.
11. M. Saleem, S. S. Padmanabhuni, A.-C. N. Ngomo, J. S. Almeida, S. Decker, and H. F. Deus. Linked cancer genome atlas database. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 129–134. ACM, 2013.
12. J. Xuan, Y. Yu, T. Qing, L. Guo, and L. Shi. Next-generation sequencing in the clinic: promises and challenges. *Cancer letters*, 340(2):284–295, 2013.
13. A. Zappa, A. Splendiani, and P. Romano. Towards linked open gene mutations data. *BMC bioinformatics*, 13(Suppl 4):S7, 2012.