Using GMQL-Web for Querying, Downloading and Integrating Public with Private Genomic Datasets

Conference Paper · September 2019

DOI: 10.1145/3307339.3343466

6 authors, including:

Marzia Settino
Universita' degli Studi "Magna Græcia" di Catanzaro
6 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Anna Bernasconi
Politecnico di Milano
16 PUBLICATIONS **43** CITATIONS
SEE PROFILE

Gaia Ceddia
Politecnico di Milano
5 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Marco Masseroli
Politecnico di Milano
193 PUBLICATIONS **1,776** CITATIONS
SEE PROFILE

Using GMQL-Web for Querying, Downloading and Integrating Public with Private Genomic Datasets

Marzia Settino
Departimento di Scienze Mediche e Chirurgiche
Universitá Magna Graecia
Catanzaro, Italy
marzia.settino@studenti.unicz.it

Anna Bernasconi
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
anna.bernasconi@polimi.it

Gaia Ceddia
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
gaia.ceddia@polimi.it

Giuseppe Agapito
Departimento di Scienze Mediche e Chirurgiche
Universitá Magna Graecia
Catanzaro, Italy
agapito@unicz.it

Marco Masseroli
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
marco.masseroli@polimi.it

Mario Cannataro
Departimento di Scienze Mediche e Chirurgiche
Universitá Magna Graecia
Catanzaro, Italy
cannataro@unicz.it

ABSTRACT

Recent integrative analyses using data from TCGA permit GWAS investigation of the genetic variants function, providing more insight than single-platform approaches. Although there has been much progress, the integration across data sets and data types remains limited. In this work we illustrate a workflow, based on the use of GMQL-Web, for combining private cancer datasets with datasets of genomic features and biological/clinical metadata sourcing from ENCODE, Roadmap Epigenomics, TCGA, as well as annotations from GENCODE and RefSeq. GMQL-Web is a web-based interface with the goal of providing a user-friendly intuitive environment for bioinformaticians and biologists who need to query genomic processed data (including public dataset not already available in the GMQL Repository) and combine them with their private datasets. Finally, we present a case study that illustrates the workflow steps to find samples extracted from a pharmacogenomic drug metabolism multi-gene platform, i.e. the Affymetrix DMET Plus platform that contain single-nucleotide polymorphisms (SNPs) that overlap with exon regions. The DMET platform is able to identify the relationship among the patients' genomic variations and drug metabolism by detecting SNPs on genes related to drug metabolism. From the obtained result, we identify only the SNPs overlapping with genes whose expression level is above a given threshold.

CCS CONCEPTS

• **Information systems → Extraction, transformation and loading**.

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3343466

KEYWORDS

TCGA; GMQL; GWAS; SNPs; integrative analysis

ACM Reference Format:
Marzia Settino, Anna Bernasconi, Gaia Ceddia, Giuseppe Agapito, Marco Masseroli, and Mario Cannataro. 2019. Using GMQL-Web for Querying, Downloading and Integrating Public with Private Genomic Datasets. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3307339.3343466

1 INTRODUCTION

Pharmacogenomics (PGx) is the study of how genes affect a person's response to particular drugs in order to optimize tailored health care for patients. Research in PGx has grown over the past decade, evolving from a candidate-gene approach, looking for a statistical correlation between specific genetic variants and diseases, to genome-wide association studies (GWAS) — a comprehensive unbiased approach that employs markers encompassing the entire genome. Single nucleotide polymorphisms (SNPs), constituting approximately the 1% of the human genome, represent one of the most common genetic alterations studied in PGx. Large-scale resources to annotate known SNPs, including dbSNP [15] and the International HapMap Project [3], help exploring the correlations between SNPs and Adverse Drug Response (ADR), i.e. an unexpected and undesired response to a drug.

The Cancer Genome Atlas (TCGA, [18]), currently available by accessing the Genomic Data Commons portal (GDC) [7], generated a large amount of omics data and it represents a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer. The heterogeneous nature of data required multiple standardization efforts (see for example [4]) to allow researchers to easily perform integrated genomic analyses.

In the literature, many studies have proposed combined analysis of cancer genomic, metabolic, and clinical data to identify patient clusters for personalized therapy. Integrative genomics studies in their broadest interpretation are reviewed in [9]. In [17] the authors

incorporate multiple layers of omic data to understand the interactive molecular system. The specific combined use of data from TCGA and GWAS is attempted in [8].
On a more operational perspective, data analysis and processing are usually performed with high level languages and pipelines that support data extraction and other common data-driven computations typical of NGS tertiary data analysis. Among other options, in this paper we selected the GenoMetric Query Language (GMQL, [13]), a high-level domain-specific query language, which is executed in an architecture for genomic data processing defined in [12].
The main contribution of this work is to provide the researchers with a simple step-by-step workflow, based on the application of the GMQL System, for combining public datasets provided by big consortia (such as ENCODE, TCGA, GENCODE or RefSeq and in addition public dataset not already available in the GMQL Repository) with private cancer datasets (e.g. sourcing from Affymetrix DMET Plus), in order to allow the researchers to gain a better understanding of the oncogenic profiles from their own experiments and studies.
The rest of the paper is structured as follows: Section 2 presents an overview of the integration of genomic data, GWAS and DMET Platform. Section 3 presents the general outline of a data analysis workflow that considers private datasets, uploads them in GMQL-Web, combines them with other publicly available datasets, performs queries and allows downloading the corresponding results. Section 4 overviews a specific case study that applies the workflow previously described. Such example employs SNPs resulting from Affymetrix DMET platform runs, GENCODE exon annotations and TCGA gene expression data. Section 5 draws conclusions and outlines future work.

2 BACKGROUND

There exists a large volume of literature in the area of integrative genomics methodologies but significant contributions have been provided by some works (e.g. [8, 9, 17]) to the integrative approach of multi-omic data sourcing from public databases such as TCGA. However, a methodology that allows the researchers to perform an integrative genomic analysis combining these public datasets with their own private datasets is still lacking.
In this background section we provide an overview on the efforts that have been previously performed in this direction. The International HapMap Project [3] is a genome-wide database of human genetic variations for discovering the sequence variants that affect common diseases. HapMap is frequently used as the reference SNP set for GWAS. Integrative analyses using data from TCGA have leveraged GWAS to investigate the functional characterization of the genetic variants [8]. Although GWAS is a powerful tool for identifying genetic variants related to complex diseases, it has important limitations such as it is discussed below.New technologies were recently developed to capture genetic differences in drug metabolism.

2.1 Genome-Wide Association Study (GWAS)

A genome-wide association study (GWAS) is an hypothesis free method to identify associations between genetic regions (loci) and traits (including diseases). GWAS have achieved great success in identifying genetic variants related to human diseases, genetic risk factors and genetic differences in drug response. Despite its several merits, it is now recognized that GWAS under certain conditions fails to identify new susceptibility loci (most SNPs discovered by GWAS have small effects on disease susceptibility).
Moreover, the large sample size required by GWAS to achieve sufficient statistical power and the large size of the produced datasets represent a computational as well as statistical bottleneck [14]. The majority of GWAS-identified variants lie in noncoding regions making it difficult to identify the causal mechanisms between SNPs and disease. GWAS analysis often discards SNPs with a low distribution of MAFs (minor allele frequency), thus excluding rare variants that, however, may have important effects on drug response. Because of these and other limitations, most of the genetic risk variants remain undetected. A successful strategy for overcoming these limitations is based on the use of pre-defined SNPs list (e.g. list of genes involved in ADME used by DMET) to interrogate variants in genes selected on the basis of their known relevance in PGx.

2.2 The Drug Metabolism Enzymes and Transporters (DMET) Platform

The DMET (Drug Metabolism Enzymes and Transporters) is a platform developed by Affymetrix to capture in human samples the allelic variants on 231 genes that are approved by the Food and Drug Administration (FDA, USA) because of their involvement in drug absorption, distribution, metabolism and excretion (ADME) [1] [10]. The DMET platform has been designed to capture several types of markers, including copy-number variations and insertions/deletions; there is evidence of its validity in identifying additional haplotypes that were not explored previously by the HapMap Project.Unlike GWAS, DMET is usually tailored to small populations that have more variables and polygenetic traits and it can detect rare variants that have a role in drug absorption, distribution, metabolism and excretion (ADME) [2],[16].
Although DMET platform allowed to identify new polymorphisms in various ADME genes demonstrating its effectiveness in the context of drug, there is a lack of tools for the integration and analysis of DMET data [1] [6]. Indeed, existing tools generally allow only to preprocess the binary data sourcing from DMET (e.g. the Affymetrix DMET-Console) and to apply simple data analysis operations but do not allow to explore the DMET data in order to discover the relations between the SNPs and the drug response.

2.3 GenoMetric Query Language (GMQL)

In order to fill this gap, the GenoMetric Query Language (GMQL) system [13], a novel high-throughput computational software for processing big data of heterogeneous genomic features, represents a powerful tool for querying and downloading public genomic data, allowing, in addition, their integration with private datasets (i.e.datasets created by a specific user). Such environment provides portable and scalable genomic data management on powerful servers and clusters (based on Apache Spark[1]).

[1] http://spark.apache.org/

The web interface uses the Web Service REST API for queries submitting, execution monitoring and results retrieving. GMQL repository contains, in addition to other public datasets, also a unified version of TCGA cancer datasets[2], made available for querying through the GMQL-Web interface.

3 A WORKFLOW FOR INTEGRATING PRIVATE AND PUBLIC DATASETS INTO GMQL-WEB

A private dataset is an alternative to public datasets and it is created by a specific user as result of upload operations or executed GMQL queries. The workflow allows the integration of private dataset with public dataset already available in the GMQL Repository as well as external (i.e. external public dataset can be uploaded following the steps of workflow such as for private dataset). A typical workflow that integrates private and public datasets is divided in three main phases (see Figure 1): *preprocessing*, *data ingestion* and *query/download* results.

3.1 Preprocessing

Before loading private datasets, tabular data (e.g. extracted from DMET platform) needs preprocessing. This phase, the first from left in Figure 1, is necessary since most of the datasets obtained as output from various systems are noisy, incomplete and possibly inconsistent. Once prepocessed, in order to employ the dataset within GMQL-web we transform it into a compliant format, the Genomic Data Model (GDM, [11]). For GDM, a GMQL dataset is defined as a collection of samples; samples, in turn, can represent a variety of genomic data. Each sample corresponds to a pair of files, which contain:

i) **region data**, describing the physical coordinates of the genome areas;

ii) **metadata**, descriptive attributes of a sample, with its biological, clinical, and experimental properties. Regions are represented with a flat attribute-based organization, including mandatory genomic coordinates and optional region properties. Metadata instead are unstructured attribute-value pairs, to capture the high variability of clinical/specimen information present in different data sources.

3.2 Data ingestion

GMQL-Web interface allows to upload private datasets and use them together with public datasets (see central phase in Figure 1). Access to GMQL-web is allowed through the following two user profiles taking into account, in different ways, the protection of privacy and security: i) *Guest user*, that does not require the user registration. It allows a limited storage and computational power for the queries and the private dataset will be deleted after a certain period of inactivity. ii) *Authenticated user*, that requires the user registration. Once registered, the user can login with credentials to his/her reserved area. Anyway, GMQL-web does not require the use of individually identifiable data.

The "Add/Upload new dataset" feature allows the user to choose among two options for uploading private datasets files: i) the *standard file-format mode* allows to use a number of file standard formats directly supported by the system[3] (e.g. BED, NarrowPeak, BroadPeak, VCF…); ii) the *custom file-format mode* can be chosen by selecting the "Custom (GTF or tab/delimited)" option and it allows to use a user-defined format following the guidelines of Gene Transfer Format (GTF) [4] or TAB-delimited formats. In this case it is required the definition of an additional XML *Schema* file describing the structure of the dataset to upload. Using either modes, uploading a new dataset into GMQL-Web requires to specify the *Dataset name* (i.e. used as a reference to the dataset) and the *Files* (i.e. the local dataset to upload). Once uploaded, the private datasets are shown in the interface Datasets viewer under the "Private" folder.

3.3 Query and download integrated datasets

GMQL-Web interface offers the possibility to build queries in the GenoMetric Query Language (GMQL, [13]). A GMQL program (or query) is a sequence of *operations* applied on one or more datasets (or variables), which results in the creation of new datasets. GMQL operators usually work on both metadata and regions, and may accept one or two operand datasets. An operation can be declared with the following structure:

```
<OV> = OPERATOR(<params>) <IV1> <IV2>
```

where OV stands for output variable, an operator can be tuned using optional parametes, and IV1 and IV2 are input variables (the first is always required, while the second one can be optional, depending on the operator). Predicates on region data use attributes from the XML schema, that is shared across samples in the same dataset. Predicates on metadata may use arbitrary attributes.

Detailed GMQL language documentation including the description of the basic operators of the language, the instructions for using them and some biological query examples can be found on GeCo Web Site [5]. Among all unary operators (i.e. SELECT, MATERIALIZE, PROJECT, EXTEND, ORDER, GROUP) and binary ones (i.e. MERGE, UNION, DIFFERENCE, MAP, COVER, JOIN) of GMQL, we here focus only on three, since they are used in the following for the example use case:

- **SELECT**. Creates a new dataset from an existing one by extracting a subset of samples from the input dataset. Conditions can be combined using boolean operators.
- **MATERIALIZE**. It writes the content of a dataset to a file, whose name can be specified, and registers the saved dataset in the repository to make it usable in other queries.
- **MAP**. It applies to two datasets, a *reference* and an *experiment*. For each sample in the experiment dataset, it computes aggregates over the values of the regions that intersect with at least one region, in at least one reference sample.The *count* aggregate counts the number of experiment regions that intersect a specific reference region.

As highlighted in the right part of Figure 1, GMQL queries can be applied to public datasets optionally combined with private ones (derived from the previous preprocessing and ingestion phases). A query is first produced in the Query editor, then compiled and

[2]This standardized, extended and integrated version is described in the OpenGDC webpage http://bioinf.iasi.cnr.it/opengdc/.

[3]https://github.com/DEIB-GECO/GMQL-WEB/wiki/file_formats

[4]http://mblab.wustl.edu/GTF22.html

[5]Data-Driven Genomic Computing http://www.bioinformatics.deib.polimi.it/geco/?try

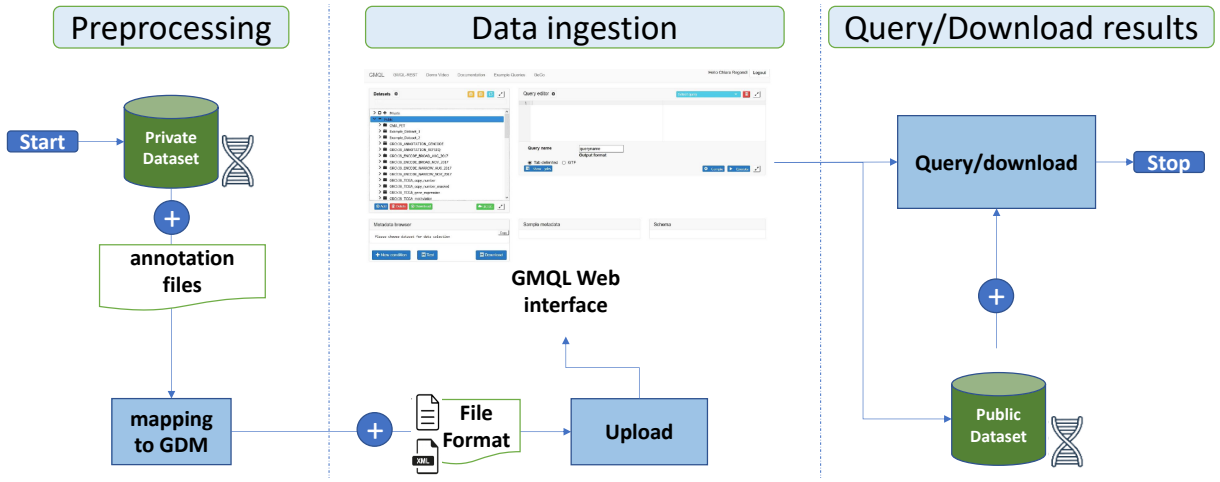

Figure 1: General workflow for integrating private and public datasets using GMQL-Web interface.

executed. The corresponding running jobs can be monitored using an ad-hoc functionality. Once results are produced (i.e. materialized) they become available in the Datasets viewer for browsing and further processing. Additionally, result datasets can be downloaded as a zip file.

4 CASE STUDY: INTEGRATING A DMET PHARMACOGENOMIC DATASET INTO GMQL-WEB

In this Section we describe a specific example workflow which combine private dataset, derived from the DMET Affymetrix platform with public datasets of annotations (from GENCODE) and gene expression levels (from TCGA). GENCODE is a scientific project which produces high-quality reference gene annotations and experimental validation for human and mouse genomes [5]. Within GENCODE, a comprehensive set of annotations of gene features was created. These include, among others: genes, transcripts, exons, protein-coding and non-coding loci, variants.

4.1 Preprocessing

Affymetrix DMET-Console allows to preprocess the raw data file generated by the Affymetrix DMET for building a comprehensive table containing, for each probe and for each sample, the detected SNP or a NoCall value (i.e. ambiguous nucleotide in the SNP). The DMET-Console output is a tab-delimited file structured as a matrix with 1936 rows (probes) and a number of columns related to the number of subjects enrolled in the analysis. Thus, the value contained in the *(i-th,j-th)* cell is the SNP detected in the *i-th* probe and belonging to the *j-th* subject.

Once the preprocessing phase is completed, the tab-delimited file obtained from DMET Console needs be transformed into a GDM compliant format. To do this, we use the DMET annotation file[6] in order to obtain the genomic coordinates for each Affymetrix probe and to map them on the genomic coordinates of the probes in DMET output (currently this is done in a manual way).

4.2 Data ingestion

As explained in Section 3.2, the user can choose among two options for uploading private datasets files, the *standard file-format mode* and the *custom file-format mode*. We used the latter mode, with the tab-delimited file format. We created a tab-delimited file, such as the one illustrated in Figure 2 (left side), containing the DMET dataset where the columns are defined as follows: the first contains the Affymetrix Probe identifier, as reported in the DMET annotation file; the next five contain the values describing the genomic regions that are characterized by their coordinates; the seventh and eighth columns contain respectively the dbSNP identifiers that map the targeted genomic region and the associated identifiers in the PharmGKB database. The remaining columns contain the detected SNPs for each subject. Before loading the DMET dataset as before defined, the XML schema file describing the structure of the dataset must be created. The right side of the Figure 2 illustrates the XML schema file used in this example. Once the dataset file and its schema are obtained, we uploaded them in GMQL-Web using the feature "Add/Upload" following the procedure described in Section 3.2.

4.3 Example queries

Here, we propose two example queries able to integrate the private DMET dataset, previously described and uploaded in GMQL-Web, with public datasets (Listings 1 and Listings 2). The first query aims to:

- select SNPs from DMET Dataset;
- select exon regions from GENCODE annotation dataset (version 27 corresponds to a GRCh38 aligned annotation effort released in August 2017);
- map exon regions on SNPs regions;
- select only SNPs regions that overlap at least one exon region.

[6] https://www.affymetrix.com/support/developer/powertools/changelog/VIGNETTE-DMET-genotyping.html

Listing 1: Query example 1 for integrating private and public datasets

```
EXON = SELECT(annotation_type == "exon" AND release_version == "27") GRCh38_ANNOTATION_GENCODE;
DMET_SNP = SELECT() DMET_SNP_uploaded;
EXON_Map = MAP(count_name: SNP_count) DMET_SNP EXON;
SNP_SELECTED = SELECT(region:SNP_count>0) EXON_Map;
MATERIALIZE SNP_SELECTED INTO SNP_Results;
```

Listing 2: Query example 2 for integrating private and public datasets

```
EXPRESSED_GENE = SELECT(gdc__project__disease_type == "Breast Invasive Carcinoma"; region: fpkm > 3.0)
                                                    GRCh38_TCGA_gene_expression_2018_12;
SNP_EXPR_Map = MAP(count_name:EXPR_count) SNP_SELECTED EXPRESSED_GENE;
SNP_EXPR_SELECTED= SELECT(region:EXPR_count>0) SNP_EXPR_Map;
MATERIALIZE SNP_EXPR_SELECTED INTO SNP_EXPRESSED_Results;
```

AM_10003	chr10	CHST3	74424904	74424904	+	rs4148943	PA26503	C/T	C/T
AM_10004	chr10	CHST3	74424940	74424940	+	rs4148944	PA26503	G/G	G/G
AM_10005	chr10	CHST3	74424987	74424987	+	rs4148945	PA26503	C/T	C/T
AM_10006	chr10	CHST3	74425470	74425470	+	rs4148946	PA26503	T/T	C/T
AM_10008	chr10	CHST3	74425752	74425752	+	rs7072219	PA26503	A/A	A/A
AM_10010	chr10	CHST3	74426048	74426048	+	rs4148949	PA26503	C/C	C/T
AM_10011	chr10	CHST3	74427103	74427103	+	rs4148950	PA26503	A/G	A/G
AM_10012	chr10	CHST3	74427181	74427181	+	rs4148951	PA26503	G/G	G/G
AM_10013	chr10	CHST3	74427411	74427411	+	rs1871450	PA26503	A/G	A/G
AM_10946	chr16	ABCC1	16270169	16270169	+	rs8056298	PA244	G/G	G/G
AM_10959	chr16	ABCC6	16285245	16285245	-	rs2238472	PA58	G/G	A/A
AM_10964	chr16	ABCC6	16297309	16297309	-	rs2856585	PA58	C/C	C/C
AM_10966	chr16	ABCC6	16305003	16305003	-	rs6416668	PA58	G/G	G/G
AM_10968	chr16	ABCC6	16306316	16306316	-	rs7500834	PA58	G/G	G/G
AM_10976	chr16	ABCC6	16310383	16310383	-	rs4341770	PA58	T/T	T/T
AM_10977	chr16	ABCC6	16312509	16312509	-	rs8058694	PA58	A/A	A/A
AM_10978	chr16	ABCC6	16312515	16312515	-	rs8058696	PA58	G/G	G/G

```
<?xml version="1.0" encoding="UTF-8"?>
- <gmqlSchemaCollection xmlns="http://genomic.elet.polimi.it/entities"
  name="biological_example_20190416_094759_MUTATION_GENE_top">
    - <gmqlSchema coordinate_system="default" type="Peak">
        <field type="STRING">probe</field>
        <field type="STRING">chr</field>
        <field type="STRING">gene_symbol</field>
        <field type="STRING">left</field>
        <field type="STRING">right</field>
        <field type="STRING">strand</field>
        <field type="STRING">DB_SNP_ID</field>
        <field type="STRING">PharmGKB Code</field>
        <field type="STRING">s1</field>
        <field type="STRING">s2</field>
        <field type="STRING">s3</field>
        .....
        <field type="STRING">s26</field>
        <field type="STRING">s27</field>
        <field type="STRING">s28</field>
      </gmqlSchema>
</gmqlSchemaCollection>
```

Figure 2: DMET output in GDM compliant format (left) and the associated XML Schema file (right).

id_sample	chr	start	stop	strand	probe	gene_symbol	DB_SNP_ID	PharmGKB Code	s1	s5	s6	s7	s8	s20	SNP_count
S_00000.gdm	chr16	16270169	16270169	+	AM_10946	ABCC1	rs8056298	PA244	G/G	G/G	G/G	G/G	G/G	G/G	5.0

Figure 3: Results of Listing 1

The input file (resulting from DMET) contains the SNPs (equipped with genomic coordinates and other information) detected by 33 probes on 28 subjects. The result of the MAP operation (EXON_Map) is a dataset which contains SNPs equipped with the count (SNP_count) of exons overlapping with that nucleotide base. Many of the SNPs do not overlap with exon, thus the SNP_count is zero. The select statement filters out samples containing a region with null count. Thus, the SNP_SELECTED variable, and the resulting materialized dataset, contains 5 SNPs (i.e. 5 SNPs are found in the exon regions). Results of this first query (see Figure 3) contain SNPs region coordinates (chromosome, start and end positions, strand) and additional characterizing attributes (gene symbol, dbSNP ID, PharmGKB ID, all SNPs detected by DMET in that region and the number of overlapped exon regions).

The query shown in Listing 2 uses TCGA Breast Invasive Carcinoma (BRCA) expression data and SNPs resulting from the previous query. The goal is to find SNPs that overlap with genes whose fpkm [7] expression level is greater than 3. The query outline can be summarized as follows:

- select Breast Invasive Carcinoma (BRCA) gene expression data with fpkm > 3;
- map the result on SNP regions (output of the query in Listing 1);
- select only SNP regions that overlap at least one highly expressed gene in BRCA dataset.

The result of the MAP operation (SNP_EXPR_Map) is a dataset that contains SNPs equipped with the count (EXPR_count) of highly expressed genes overlapping with that nucleotide base. Many of the SNPs do not overlap with genes, thus the SNP_count is zero. We filter such cases with a SELECT statement as it results in the materialized dataset.

Results of the second query (see Figure 4) contain SNPs region coordinates and additional characterizing attributes structured as

[7] fpkm:Fragments Per Kilobase of transcript per Million mapped reads

id_sample	chr	start	stop	strand	probe	gene_symbol	DB_SNP_ID	PharmGKB Code	s1	s2	s3	s4	s5	SNP_count	EXPR_count
S_00000.gdm	chr16	16270169	16270169	+	AM_10946	ABCC1	rs8056298	PA244	G/G	G/G	G/G	G/G	G/G	5.0	1.0
S_00001.gdm	chr16	16270169	16270169	+	AM_10946	ABCC1	rs8056298	PA244	G/G	G/G	G/G	G/G	G/G	5.0	1.0
S_00002.gdm	chr16	16270169	16270169	+	AM_10946	ABCC1	rs8056298	PA244	G/G	G/G	G/G	G/G	G/G	5.0	1.0
S_00003.gdm	chr16	16270169	16270169	+	AM_10946	ABCC1	rs8056298	PA244	G/G	G/G	G/G	G/G	G/G	5.0	1.0
S_00004.gdm	chr16	16270169	16270169	+	AM_10946	ABCC1	rs8056298	PA244	G/G	G/G	G/G	G/G	G/G	5.0	1.0
S_00005.gdm	chr16	16270169	16270169	+	AM_10946	ABCC1	rs8056298	PA244	G/G	G/G	G/G	G/G	G/G	5.0	1.0

Figure 4: Results of Listing 2

the results of the first query. However, in this case results have one more attribute called EXPR_count derived from the overlapping between SNP_SELECTED and EXPRESSED_GENE. MAP operation can be computationally expensive since it performs a cartesian product between the reference and the experiment dataset samples. However, in our specific case, the employed input datasets do not have critical sizes and are handled in acceptable computation time. In particular, the entire execution of the Listening 1 and Listening 2 required 24 minutes and 02 seconds.

5 CONCLUSION

Integrative genomics methodologies provide a way of dealing with huge and heterogeneous biological data to facilitate a better understanding of the cellular biology systems. However, a methodology that allows the researchers to perform an integrative genomic analysis combining public datasets with their own private datasets is still lacking. The main contribution of this work is a step-by-step workflow for combining public datasets provided by big consortia (such as GENCODE and TCGA) with private cancer datasets in order to allow the researchers to gain a better understanding of the oncogenic profiles from their own experiments and studies. The proposed workflow is based on the use of the GMQL-Web, a web-based interface with the goal of providing a user-friendly intuitive environment for bioinformaticians and biologists who need to query public genomic database and combine them with their private datasets. A case of study that applies the workflow on datasets resulting from Affymetrix DMET platform experiments and public genomic data was illustrated. Future work will regard i) the development of a tool for generating the GDM compliant format of DMET output; ii) the integration of DMET datasets with PharmaGKB (Pharmacogenomics Knowledge Base) in order to obtain additional pharmacogemomics information.

ACKNOWLEDGMENTS

We wish to thank P. Tagliaferri, P. Tassone, M.T. Di Martino, M. Arbitrio and F. Scionti for introducing us to the DMET platform.

REFERENCES

[1] G. Agapito, P. H. Guzzi, and M. Cannataro. 2015. DMET-Miner: Efficient discovery of association rules from pharmacogenomic data. *J Biomed Inform* 56 (Aug 2015), 273–283.

[2] M. Arbitrio, M. T. Di Martino, F. Scionti, G. Agapito, P. H. Guzzi, M. Cannataro, P. Tassone, and P. Tagliaferri. 2016. DMET (Drug Metabolism Enzymes and Transporters): a pharmacogenomic platform for precision medicine. *Oncotarget* 7, 33 (08 2016), 54028–54050.

[3] INTERNATIONAL HAPMAP CONSORTIUM. 2003. The international HapMap project. *Nature* 426 (2003), 789–796. Issue 6968.

[4] Fabio Cumbo, Giulia Fiscon, Stefano Ceri, Marco Masseroli, and Emanuel Weitschek. 2017. TCGA2BED: extracting, extending, integrating, and querying The Cancer Genome Atlas. *BMC bioinformatics* 18, 1 (2017), 6.

[5] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. 2018. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* 47, D1 (2018), D766–D773.

[6] P. H. Guzzi, G. Agapito, M. T. Di Martino, M. Arbitrio, P. Tassone, P. Tagliaferri, and M. Cannataro. 2012. DMET-analyzer: automatic analysis of Affymetrix DMET data. *BMC Bioinformatics* 13 (Oct 2012), 258.

[7] Mark A Jensen et al. 2017. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* 130, 4 (2017), 453–459.

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

[9] V. N. Kristensen, O. C. Lingj?rde, H. G. Russnes, H. K. Vollan, A. Frigessi, and A. L. B?rresen-Dale. 2014. Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 5 (May 2014), 299–313.

[10] Jing Li, Luyong Zhang, Hang Zhou, Mark Stoneking, and Kun Tang. 2010. Global patterns of genetic diversity and signals of natural selection for human ADME genes. *Human Molecular Genetics* 20, 3 (11 2010), 528–540. https://doi.org/10.1093/hmg/ddq498 arXiv:http://oup.prod.sis.lan/hmg/article-pdf/20/3/528/17254605/ddq498.pdf

[11] Marco Masseroli et al. 2016. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods* 111 (2016), 3–11.

[12] Marco Masseroli, Arif Canakoglu, Pietro Pinoli, Abdulrahman Kaitoua, Andrea Gulino, Olha Horlova, Luca Nanni, Anna Bernasconi, Stefano Perna, Eirini Stamoulakatou, et al. 2018. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. *Bioinformatics* 35, 5 (2018), 729–736.

[13] M. Masseroli, P. Pinoli, F. Venco, A. Kaitoua, V. Jalili, F. Palluzzi, H. Muller, and S. Ceri. 2015. GenoMetric Query Language: a novel approach to large-scale genomic data management. *Bioinformatics* 31, 12 (Jun 2015), 1881–1888.

[14] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81, 3 (2007), 559 – 575. https://doi.org/10.1086/519795

[15] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29, 1 (01 2001), 308–311. https://doi.org/10.1093/nar/29.1.308 arXiv:http://oup.prod.sis.lan/nar/article-pdf/29/1/308/9905801/290308.pdf

[16] Tristan M Sissung, Bevin C English, David Venzon, William D Figg, and John F Deeken. 2010. Clinical pharmacology and pharmacogenetics in a genomics era: the DMET platform. *Pharmacogenomics* 11, 1 (January 2010), 89âĂŤ103. https://doi.org/10.2217/pgs.09.154

[17] Yan V. Sun and Yi-Juan Hu. 2016. Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. *Advances in Genetics* 93 (2016), 147 – 190.

[18] K. Tomczak, P. Czerwi?ska, and M. Wiznerowicz. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 19, 1A (2015), 68–77.