M.S. THESIS

BioVLAB-TCGA: Web-based TCGA Omics data Mapping on KEGG Pathway System

BioVLAB-TCGA: 웹 기반 TCGA 오믹스 데이터 KEGG 패스웨이 매핑 시스템

FEBRUARY 2017

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Choi Saemi

M.S. THESIS

BioVLAB-TCGA: Web-based TCGA Omics data Mapping on KEGG Pathway System

BioVLAB-TCGA: 웹 기반 TCGA 오믹스 데이터 KEGG 패스웨이 매핑 시스템

FEBRUARY 2017

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Choi Saemi

Abstract

BioVLAB-TCGA: Web-based TCGA Omics data Mapping on KEGG Pathway System

Choi Saemi
Department of Computer Science and Engineering
College of Engineering
Seoul National University

TCGA, The Cancer Genome Atlas, is a database which provides omics data to public. National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) in the United States had generated this comprehensive, multi-dimensional genomic database. Even though many scientists have tried TCGA omics data analysis, selecting target genes or patients always depends on their prior biological knowledge of their own. To enhance TCGA data usability, there needs to be a system which provides biological filtering and visualizing experimental environment. BioVLAB-TCGA system has fully implemented these requirements. In the system, graphical pathway maps from Kyoto Encyclopedia of Genes and Genomes (KEGG) were used for gene selection and visualization. KEGG Pathways are fully biological meaning because they were drawn by human with biology literatures. Likewise, TCGA clinical data and PAM50 classification were applied for patients selection. Once scientists simply click the pathway and patient clinical option on web

pages, then the web front-end creates URL dynamically and requests data onto REST API. Web front-end code reads the result and visualize them as figures. KEGG pathway entries are colored after grading. while venn diagram and OncoPrint describes the landscape of selected genes and patients. With this system, scientists can be given an insight on biological meaning of selected TCGA data.

Keywords: The Cancer Genome Atlas, Breast Cancer, KEGG Pathway, Omic Visualization, Gene expression analysis, RNA sequencing, Mutation, Copy number variation, TF-TG Correlation, Data mining.

Student Number: 2015-21270

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Cancer Genome Atlas (TCGA)

1.1.1 TCGA Omics Data

As the era of omics data begins, there need to have more databases and analytic tools in the world. Omics data is big data collected from biological or medical experiment. This suggests combing different genomics, proteomics, and transcriptomics data (Somerville and Somerville, 1999). Some institutions or universities such as National Center for Biotechnology Information (NCBI), Broad Institute, Kyoto University, and et al have provided omics databases and analytical tools such as The Cancer Genome Atlas (TCGA), Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway, Broad GDAC Firehose, cBioPortal and so on. In this study, we focused on the TCGA Breast Cancer database and function based human-curated graphical gene maps KEGG pathway database.

TCGA has 33 different projects such as, Adrenocortical Carcinoma

(ACC), Bladder Urothelial Carcinoma (BLCA), Brain Lower Grade Glioma (LGG), and Breast Invasive Carcinoma (BRCA) etc. taken from doctors or biologists. In TCGA database, 2.5 petabytes of data totally were stored from 20 collaborating institutions. The number of papers which used TCGA data has increased exponentially. It counted only 9 papers in the year of 2005, when TCGA launched. However, papers have been published journals used TCGA counted to 700 in 2016. And it amounts to 2,000 papers stored in PubMed totally.

1.1.2 TCGA-BRCA

TCGA-BRCA is one of the major projects in TCGA. Among 11,000 patients registed for TCGA, 1,098 breast cancer patients have been counted currently. Breast Cancer is a cancer developed in breast (Network et al.,2003). In 2012, it has been found with 1.68 million cases and 522,000 deaths (Stewart et al., 2016). Because of the significance of breast cancer study, TCGA-BRCA data have been used widely. Researches that used TCGA-BRCA data directly have been reported. Especially, studies have found loci or specific gene candidates with large scale TCGA-BRCA data were active (Michailidou et al., 2013; Li et al., 2013). In addition to this, studies with breast cancer subtype related genes were also reported (Bose et al., 2013; Creighton et al., 2012; Prat et al., 2013). And patient survival or therapeutic roadmap studies also published (Van De Vijver et al.,2002; Ellis and Perou, 2013). TCGA-BRCA omics data consists of gene expression, mutation, and copy number variation.

1.2 Kyoto Encyclopedia of Genes and Genomes Pathway

KEGG, which stands for Kyoto Encyclopedia of Genes and Genomes, have collected graphical pathway maps, orthologous group tables, and molecular catalogs. Pathway is a map that contains gene entries and their edges. This describes biological signaling or gene functions. Now, 308 pathways are stored in KEGG pathway database. KEGG pathway database gives knowledge base of gene functions, linking genomic information about higher order functional information. And graphical representations of cellular processes, metabolism, membrane transport, signal transduction and cell cycle (Kanehisa and Goto, 2000) were given by KEGG pathway. In this study, to test system performance, ErbB signaling pathway will be navigated and significantly dealt with. *ERBB2* is a gene that affects breast cancer and its signaling pathway is a biological pathway that explains how the *ERBB2* gene works in the human body.

Gene Expression		Mutation		Copy Number Variation		Clinical	
patient	1,093	patient	977	patient	1,080	patient	1,098
gene	20,531	gene	17,279	gene	24,776	attribute	21
total	22,440 ,383	total	86,765	total	26,758 ,080	total	1,098

Table 1 BioVLAB-TCGA stored TCGA data status.

1.3 Motivation

1.3.1 TCGA Data on KEGG Pathway

Scientists have used multi-omics data to identify the biological mechanisms. Mapping gene expression data onto KEGG pathway is one of the widely used approaching to analyze and find biological meanings. There have been studies for automatic mapping program or system. Pathview (Luo and Brouwer, 2013) is a R/Bioconductor package, which mapped user data onto visualized pathway. Pathview downloads the pathway graph, maps and integrates user data onto the pathway. Finally it renders pathway graphs with the user-mapped data. Pathview is helpful but its environment is still static. Scientists should download packages and install them. Pathway Inspector is a web application helping scientists to find patterns of gene expression given from complex RNAseq custom experiments (Bianco et al., 2016). Pathway Inspector combines identification of differentially expressed genes (DEGs) and builds topology-based analysis of enriched pathways.

On the other hand, TCGA, has many multi-omics analytic tools

reported. Integrative Genomics Viewer (IGV) (Robinson et al., 2011), cBioPortal for Cancer Genomics (Gao et al., 2013), and Broad GDAC Firehose were widely used nowadays.

Even though both TCGA and KEGG pathway based studies are active these days, their concepts does not reflect each other. KEGG pathway based analysis would be done with custom data provided by scientist. TCGA researches tried a lot of work but they need more dynamic analysis tools because pathway-based approaches help preventing scientists from depending on their prior biological knowledge. Still, some of these tools are not based on web. Therefore, a web-based system for mapping TCGA multi-omics data on KEGG pathways is inevitable.

1.3.2 Our Work

In this study, we designed and implemented a web-based system called BioVLAB-TCGA. This biological virtual lab using TCGA data onto web space lets scientists achieve a discovery while narrowing down genes with respect to KEGG pathways and TCGA patient options. This system supports mapping TCGA data onto KEGG pathway by acquiring data from modeled and designed database. In web front-end, built-in code tries to make decisions on TCGA omics data. For instance, some questions like what entries, a gene group, are up regulated or down regulated. Furthermore, scientists can get information about candidate entries in case of up regulating, mutation one or more, and the Copy Number Variation (CNV) amplified more

than 30%.

Visualization is also major function of BioVLAB-TCGA. Scientists can achieve not only TCGA data onto KEGG pathway, but also pathways summary, OncoPrint, a patients-mutation and CNV map, and data table by simply clicking on web page. Therefore scientists could have struck a great biological idea by taking a glimpse of brief visualization.

Chapter 2

Materials and Methods

2.1 System Structure

BioVLAB-TCGA consists of three parts - Database system, REST API, and Web front-end. As a first technical base, Database system which structured omics data such as TCGA-BRCA gene expression RNASequencing normalized counts, mutation, CNV data, clinical data, and KEGG pathway data. And its management system is MySQL. Second part of BioVLAB-TCGA is REST API (application programming interface). This interface given by Django web framework provides structured omics data from database in response to client's requests. The last part of BioVLAB-TCGA is the front-end. This part calls TCGA-BRCA omics data with ajax, one of the useful asynchronous communication method while dealing with

various repositories for biology databases (Aravindhan et al., 2009). Once the web browser which is in side of clients requests data through URL, REST API would send data after throwing a query to database system and returned query result (queryset) by database virtual view. Clients could approach those data by not only BioVLAB-TCGA web service, but also web browser or terminal.

BioVLAB-TCGA is developed on the basis of this database system, REST API, and web front-end. BioVLAB-TCGA is fully navigating TCGA omics data and KEGG pathway data through REST API by aggregating and visualizing those data for each web page given client's requests. In summary, BioVLAB-TCGA main technology stack is given below.

- Database: MySQL
- REST API: Django web framework
- Web front-end: Ajax, jquery, javascript, html, css et al.

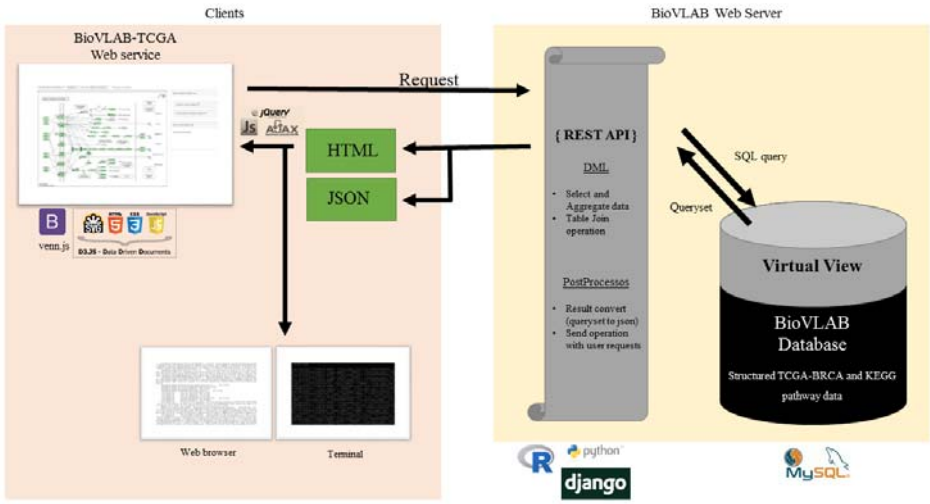


Figure 2.1 System Structure for BioVLAB-TCGA.

BioVLAB-TCGA system structure overview. The right part describes BioVLAB server side while the left part is about client server sided. BioVLAB server builds the database which contains structured TCGA and KEGG omics data. Text data have been abstracted with MySQL in which tables imply the omics data while each tuple implying each data itself. For example, ' KEGGPathway' table contains 308 pathways information its tuples. Furthermore, database is abstracted one more time by virtual view that simplifies tables with ' join' functions. REST API could get the dataset when it throws query to database. These datasets will be provided as sort of types by Django framework so that the client accesses to data by terminal and web browser.

2.2 Database Model on Relational Database

2.2.1 Data Information

Gene expression data are described as gene-level RNA-seq expression data especially on normalized RSEM value. They were obtained from the standardized analysis-ready TCGA data, Broad GDAC Firehose (http://doi.org/10.7908/C18W3CNQ) stddata__2015_08_21 run (Onitilo et al., 2009) breast cancer subtypes based on er/pr/her2 expression (Trichopoulos et al., 1972; Bovolenta et al., 2012; Liberzon et al., 2011; Network et al., 2012; Bianco et al., 2016; Luo and Brouwer, 2013). Gene expression data were given by 20,531 genes for 1,093 patients.

Mutation data provide variant classification, variant type, genome change, and protein change etc. Data were given by 17,279 genes for 977 patients. Because not every patient has mutation on every gene, the number of mutation data is less than gene expression data.

CNV data describe how much amplified, deleted or neutral the specific gene is. The data were marked as an integer ranging from -2 to 2. The minus means deletion and the plus means amplification while zero is neutral. This data were given by a method called GISTIC2 (Mermel et al., 2011) which is a sensitive and powerful way to identify genes targeted by somatic copy number alterations (SCNA). SCNA is known to drive cancer growth. CNV data were given by 24,776 genes for 1,080 patients.

For clinical data, some attributes which are used very frequently were selected for BioVLAB-TCGA. Studies used age, vital status,

pathology stage, gender, er/pr status information to analyze TCGA-BRCA data (Azim et al., 2015; Benevolenskaya et al., 2016). 1,098 patients were recorded in TCGA-BRCA. In addition to this, BioVLAB-TCGA supports PAM50 classification, a method that uses 50 genes for classifying breast cancer, to provide breast cancer subtype such as basal, luminal A or B, her2 and so on.

Patient barcode for each data is given by a sequence of string as the same representation to TCGA. For example, 'TCGA-02-0001-01' is kind of a patient sample. It means 'Project-TSS-Participant -Sample'. This sample was registered for TCGA project and it's brain tumor GBM (02) sample from MD Anderson. And the patient is the first participant for GBM study (0001) while the sample type was solid (01). In case of gene, gene unique entrezID and symbols were stored in database.

2.2.2 TCGA-BRCA Data Modeling

In this study, a relational database was built to implement BioVLAB-TCGA because TCGA is not dynamic database. With the database management system, MySQL, we built schema shown in Figure 2.2. Each block is matched to web framwork's model, and table in MySQL. Arrows mean the keys. By joining tables with key values, then BioVLAB-TCGA can easily achieve complicated result with scientist's dynamic demands.

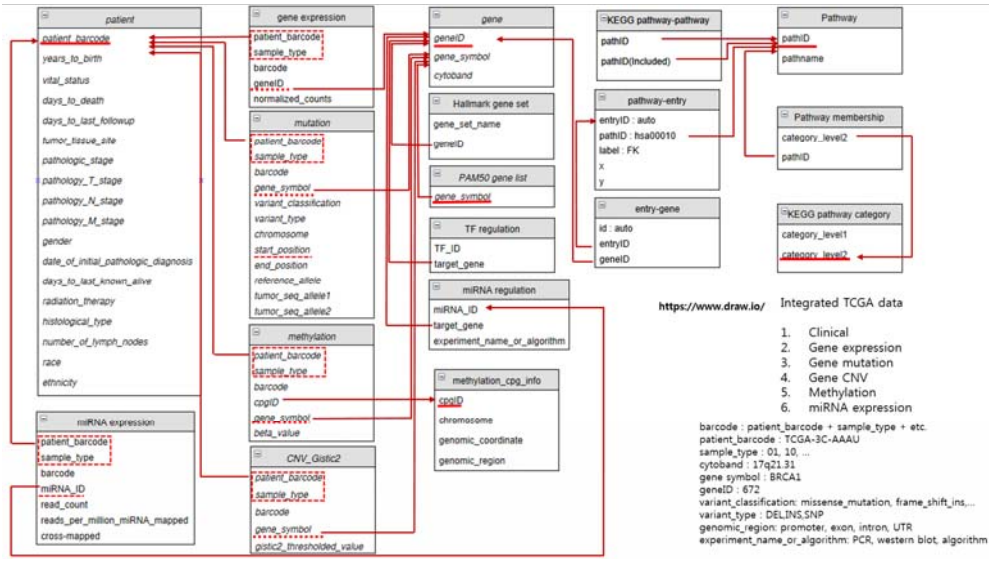


Figure 2.2 BioVLAB-TCGA Database Schema.

Each block is matched to web framework's model, and table in MySQL. Arrows mean the keys.

2.2.3 KEGG Pathway Data Modeling

BioVLAB-TCGA used KEGG pathway data by accessing KEGG API (http://www.kegg.jp/kegg/docs/keggapi.html) and KEGG PATHWAY Database (http://www.genome.jp/kegg/pathway.html). Both are provided by Kanehisa Laboratoreis in Kyoto University and University of Tokyo. The KEGG API stores relations between pathway and gene, KEGG gene ID and gene common names et al. This is RESTful API that support data access only unlike other programming interfaces. KEGG PATHWAY database contains graphical representations of many biological functions (Kanehisa and Goto, 2000).

Gene information such as gene names, KEGG IDs, relations between pathways and genes et al were acquired from KEGG REST API. Information about entries, nodes which means a group of genes on KEGG pathways, such as positions of entries in pathway, gene list in entries, and the first and second level of categories of KEGG pathways were also acquired from KEGG PATHWAY database by crawling using python program.

Structuring data from KEGG REST API and KEGG PATHWAY database, BioVLAB-TCGA has created classes which are mapped with tables in MySQL. KEGGPathway, KEGG_Entry_Gene, KEGG_Pathway_Pathway_Membership, KEGG_Category_Membership, and KEGG_Pathway_Entry_Membership classes were built on database.

2.3 REST API

Representational State Transfer (REST) Application Programming Interface (API) will be serviced by BioVLAB-TCGA for clients. REST API is an API architecture that supports an uniform interface, stateless, cacheable, client-server, layered system, and self-descriptiveness (Fielding and Taylor, 2002). A well-designed REST API is a must-have feature of today's web services (Masse, 2011). Some bioinformatics databases such as KEGG REST API (Kanehisa et al., 2004), TCGA GDC API (Grossman et al., 2016), Ensembl API (Yates et al., 2014) have been serviced and even further omics data analysis tools have provided REST APIs (Bhagat et al., 2010; Gauthier et al., 2015; Roberts et al., 2016).

As described above, KEGG and TCGA have provided REST API through system structure. Because BioVLAB-TCGA web service is helping scientists to navigate TCGA data on KEGG pathways, REST API can simply filter genes from KEGG pathway IDs at first, and aggregating TCGA data after getting query result from MySQL. Like other REST API, clients or scientists can access to data with simple endpoint coding.

Examples:

- bhi2.snu.ac.kr:8080/landscape
- bhi2.snu.ac.kr:8080/search?keyword=erbb1
- bhi2.snu.ac.kr:8080/genes/hsa00010+hsa00030
- bhi2.snu.ac.kr:8080/pathways/hsa00010/related_pathways
- bhi2.snu.ac.kr:8080/TCGA-BRCA/hsa00010/CNV?gender=male

First domain address is the server BioVLAB-TCGA is running. After the slash mark, at least one argument will be given. First argument represents the data the clients want to retrieve. 'landscape' means the status of KEGG and TCGA data status which are resided in BioVLAB server. BioVLAB will be updated periodically. While 'search' means the pathways list will be given after searching gene names or pathway names, 'genes' means the gene list in pathways given by arguments 2. Furthermore, 'pathways' returns the pathways result of argument 3's endpoint filtered by argument 2. And the last, most significant function of implementing BioVLAB-TCGA web service. It helps to grading KEGG pathway's entries in colors that were given from TCGA-BRCA omics data. The example is meaning it will provide the result of aggregating CNV data from TCGA-BRCA data given the patients option is male and the genes filtered by the pathway 'hsa00010'. For more customized use, reference Table 2.

Endpoints			Parameters	Description
arg1	arg2	arg3		
doc	-	-	-	API documentation
landscape	-	-	fiels[TCGAgenes, KEGGPathways, Pathway_Pathway _Membership,et al.]	Load KEGG pathways, TCGA-BRCA data status
search	-	-	keyword fields[pathID, pathname, labels, count]	Get search result by keyword. Gene or pathway names could be input. ex) keyword=erbb1, keyword=PI3K signaling pathway
genes	* {pathIDs}	count tftg driver	fields[geneID, gene_symbol]	Retrieve genes information. count, TFTG correlation, is Driver gene filtered by pathIDs.
pathways	* {pathIDs}	related _pathways entries summary img	fields[included_pathID, category_level2_ id, pathID, pathname, geneID, gene_symbol]	Retrieve pathways information. related_pathways, entries, summary, images filtered by pathIDs
TCGA-BRC A	* {pathIDs}	patients_list patients _count fold_change mutation CNV	view[pathID, [patients_option] fields[patient_bar code, geneID, gene_symbol, avg_normal_exp, avg_tumor_exp, selected_tumor_e xp, et al.]	Retrieve pathways and TCGA-omics information. patients_list, patients_count as patients information, and fold_change, mutation, CNV as omics data. aggregated by calculating or counting given by patients_option filtered by pathIDs.

Table 2 BioVLAB-TCGA REST API Endpoint.

It consists of three arguments and parameters. If clients input each endpoint given parameters in URL following BioVLAB-api domain address, then the system returns the dataset.

2.4 Front-End Visualization

Given REST API, the web browser resided in client´s side will be provided omics data very easily. BioVLAB-TCGA web service is an application which visualizes TCGA-BRCA data on KEGG pathway. As technical web languages - Javascript, html, and css - help to build the structure of web pages, some application packages - JQuery, Bootstrap, D3.js, venn.js, and etc. - were used to visualize things. And JQuery-AJAX is the key technique to load omics data from REST API for this service.

For example, in a very first page, a search engine-like page, the client will put some interesting words in text input-box. After requiring the result by clicking button, the service will retrieve the result of pathways which contains the gene inside the pathway's entry or the pathways themselves through REST API endpoint 'search'. After loading the pathways the client is interested in, the service is ready for waiting for any change in the page. If the client changed the patients option, then patients_count endpoint is created and JQuery-AJAX requests the data through REST API. To display each entry in pathway, the URL will be given with endpoint 'fold_change', 'mutation', or 'CNV' with parameters the client selected from control panel. Then the JQuery-AJAX also requests the omics data through REST API. Three of these major aggregating functions would be done simultaneously and then each entry will have each attribute and their values by the action of JQuery code which is also resided in front-end.

Finally, for the purpose of overview TCGA-BRCA omics data and deepening its understanding, Javascript code find and highlight some patterns of entries which the scientist may want to study. Gene expression could be seen by fold_change up, constant, and down status, and mutation is also represented by mutation count - Non exist, Exist, and repeatedly exist-. Amplification enriched or deletion enriched CNV information is also provided and visualized by color in entries. Client could specify one or three of all omics features. One entry can be filled with one color or three colors divided by three features. Javascript code makes decision what entries are the candidate for clients. If biologists who study breast cancer in human could select gene expression up, mutation exist (one or more times), and CNV amplification enriched, then select the options in highlight box below control panel. After calculating, the web page will highlight the entries satisfied the options while the others are shaded. Additionally, if TF-TG correlation are existed or that gene is driver gene, then entries which contain TF-TG correlation, and oncogene or tumor suppressor are marked and each capital word in the list of its entry information.

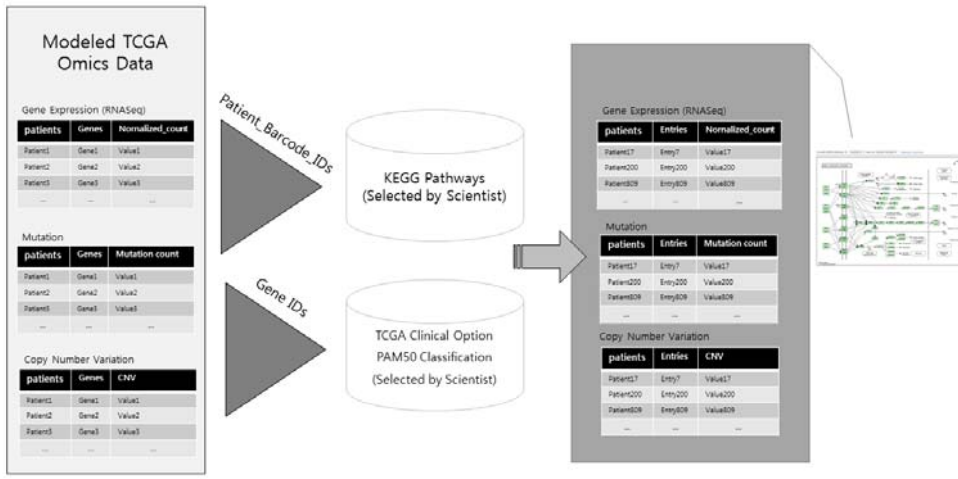


Figure 2.3 BioVLAB-TCGA Front-End Visualization.

BioVLAB-TCGA web front-end catches these data and visualizes for clients. Scientists can easily select pathways, then filter genes from TCGA data. Likewise patients also filtered after selecting patient options.

Chapter 3

Results

3.1 TCGA Visualization on KEGG pathway

Once the data acquired from the BioVLAB-TCGA server, through REST API, scripting language Javascript tries to draw entries on KEGG pathways. Each entry has attributes which represent gene expression, mutation repeatability, CNV count. Using jQUERY, change the color level of each entry.

As a result, BIoVLAB-TCGA enables scientists to navigate TCGA dynamically.

To test the system, we used Her2 breast cancer subtype as a query pathway. First of all, on the home page in BioVLAB-TCGA, put 'ERBB2' as a query. Because the system already has the gene and pathway names, the client can select the exact name with drop-down box. After clicking search button, then a list tabled come out on the page. If the client put more keywords on search input, then he or she can find more list tables as many as keywords.

The list table represents KEGG pathways which contain *ERBB2* gene. 'EGFR tyrosine kinase inhibitor resistance', 'Endocrine resistance', 'Platinum drug resistance', 'ErbB signaling pathway' and 15 more pathways have listed as a result. The client could find related pathways also. The table could be appended by clicking a pathway title. It shows the description of pathway and the related pathways which share genes with the pathway just clicked. If the client clicks 'ErbB signaling pathway' what we are really interested in, then client could select 9 related pathways below. Then the client selects the pathways might want to navigate, and click 'load pathways' button to load KEGG pathway image and patient control panel.

To see the important gene entries, try put subtype as Her2. BioVLAB will show the number of Her2 type breast cancer patients and we can see 161 patients count as a result. After clicking 'load TCGA' data again. Small circle appears on the screen to notice loading time. Approximately 3~5 seconds after, nothing has been changed on KEGG pathway but control panel for omics data selection has appended.

In case of gene expression, we used fold change values. It implemented with two different bases - tumor pool and normal pool - average values were used. And for mutation, mutations in the same position have been counted. Finally, CNV is aggregated with amplification or deletion enriched, neutral.

If the client clicks the gene expression, then easily find that *ERBB2* gene entries are up regulated while *NRG1, EGFR* gene

entries are down regulated.

After receiving omics data option, BioVLAB-TCGA provides highlight pattern function for biological insights. In case of highlighting up gene expression, one or more mutations, and amplification/both enriched, then *ERBB2, SRC, PRKCA,* and *RPS6KB1* have been highlighted. An *EGRF* gene entries have interesting pattern that mutation exists and CNV amplified while gene expression was down. Not only *ERBB2*, the other name *Her2*, is important in Her2 type breast cancer, but also *SRC* and *EGRF* are also reported as a biomarker (Moiseeva et al., 2006).

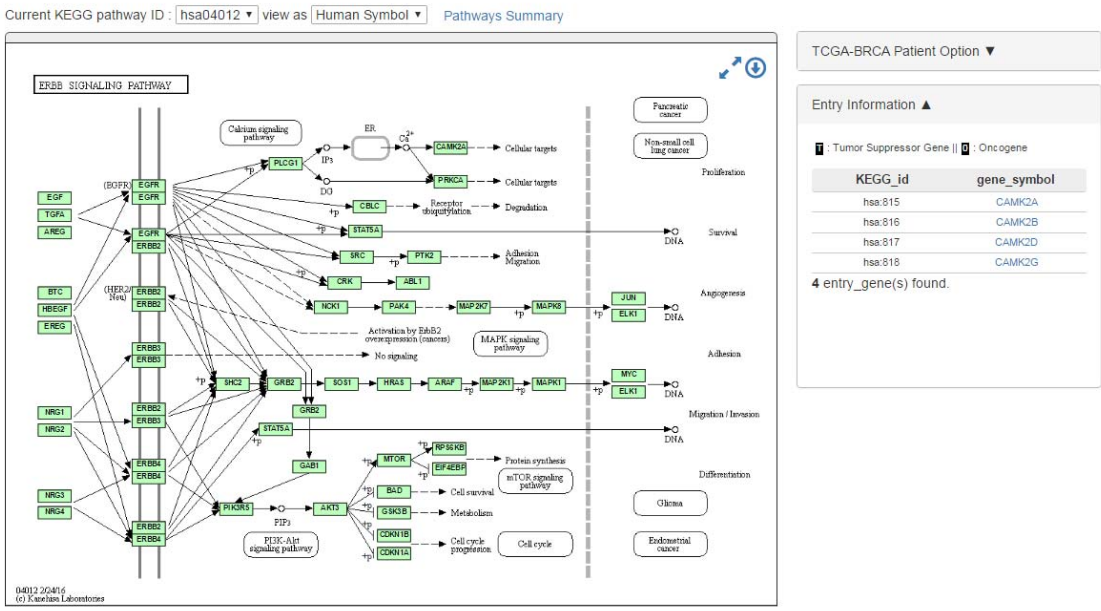


Figure 3.1 KEGG Pathway before loading TCGA data.

Green colored nodes are the entry, a group of gene, in KEGG pathway. Before loading TCGA data, only entries which contain genes found in human are in basic green color.

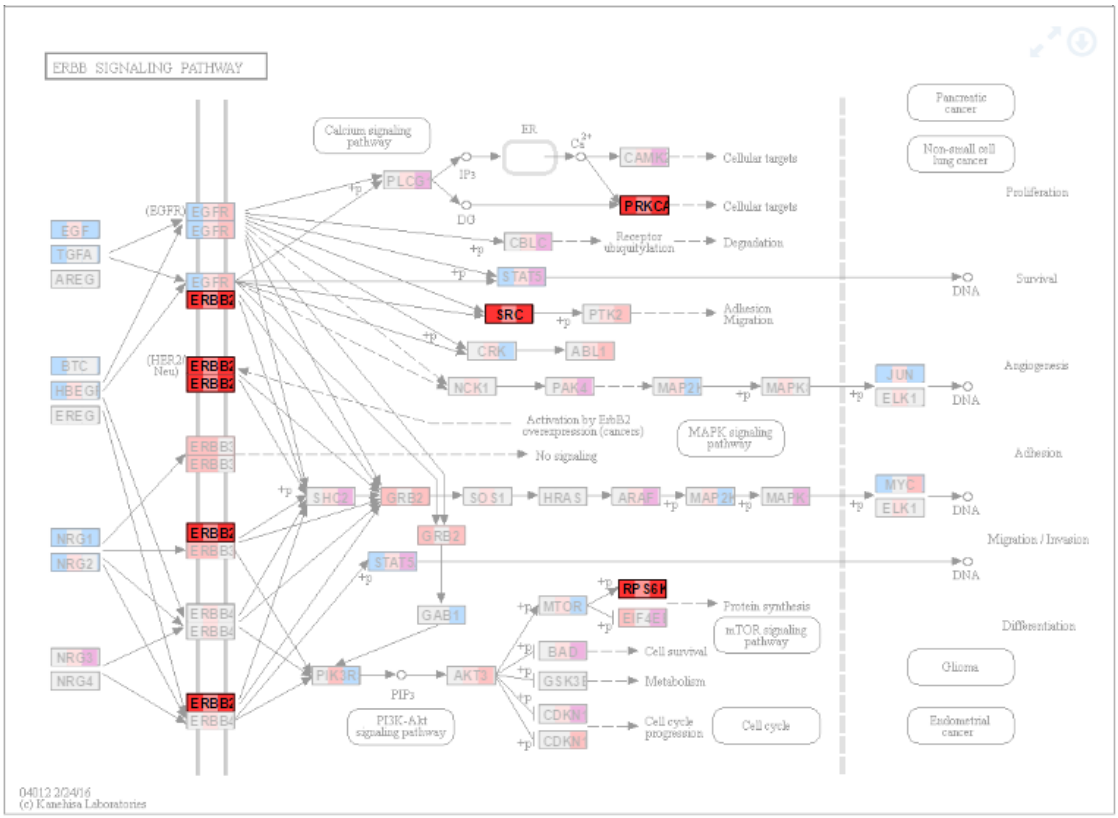


Figure 3.2 Omics Pattern Detected.

Scientists can highlight entries which are interested in their researches. In this case, in ErbB signaling pathway which is important to deal with breast cancer Her2 subtype, the entries mainly consisted of *ERBB2, SRC, PRKCA,* and *RPS6KB1* gene entries are important with respect to up gene expression, one or more mutations, and amplification enriched or deletion-amplification both enriched status.

3.2 Pathway Summary and OncoPrint View

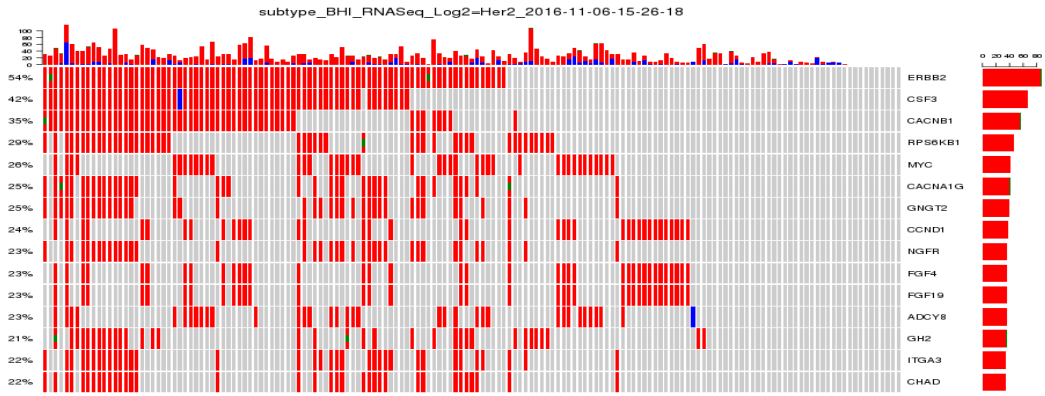
BioVLAB-TCGA web service provides in depth approaches for each omics data given patients option. Analyze omics data by fully access, data need to be provided by patients not allowing aggregating. To achieve this, server side R code is needed to visualize omics data fully. In BioVLAB-TCGA web service, TCGA-BRCA raw data and OncoPrint are provided. This will appear after highlighting some entries the client may want to study. With this process, client could narrow down the candidate of genes with respect to pathways.

OncoPrint is a main visualization figure given patients option, omics layer, and KEGG pathway. OncoPrint is a heatmap which represents the patients list on x-axis while gene lists on y-axis. Each square are gray without dot when it is not mutated and CNV is neutral. If CNV is red, that gene is amplified with those patients, otherwise if CNV is blue, that gene is deleted. Mutation is marked with green dot inside the squares. This represents the whole landscape of patients and genes various omics layer status. OncoPrint represents omics data level by color. One main method, called memo sorts which derives the result of mutually exclusively amplified or deleted gene has been used. Mutual exclusivity helps to understand the function of gene regulation.

Venn.js, OncoPrint R package could support draw venn diagram of pathways and OncoPrint to take mutations and CNVs in at a glance in specific KEGG pathways.

Finally, BioVLAB-TCGA is helping scientists by narrowing down omics data. First, selecting KEGG pathways filtered genes which are related to specific biological purpose. Second, pouring TCGA-BRCA omics data to KEGG pathway’s entries. By enabling highlighting entries which satisfied client’s interests.

(A)



(B)

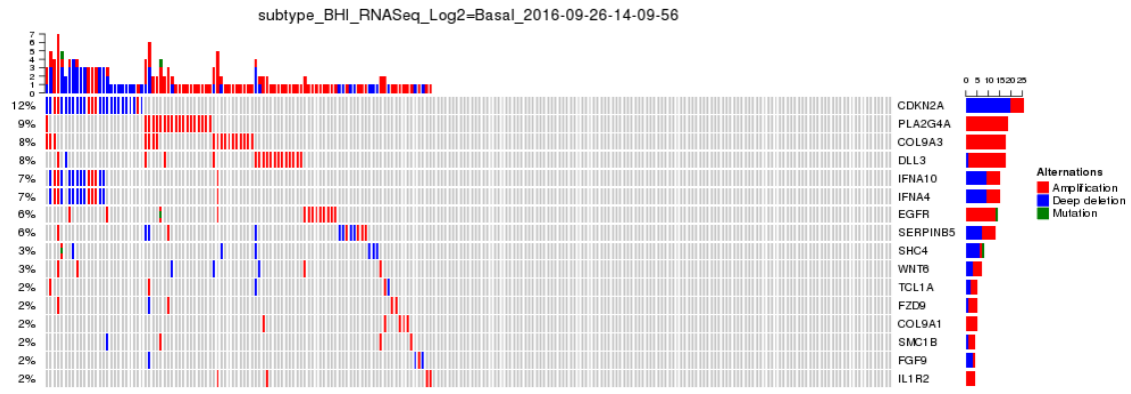


Figure 3.3 In-depth visualization of patient-CNV, mutation by OncoPrint. In ErbB signaling pathway, if Her2 type breast cancer patients were selected(A), then we can check *ERBB2* gene is highly ranked with respect to mutation and CNV. However, if Basal type selected in the same pathway(B), then *CDKN2A*, *PLA2G4A* and other genes are highly ranked. However, basal type breast cancer is more mutually exclusive so that it has more omics heterogeneity.

Chapter 4

Discussion

The service enables access to omics data together. Gene expression, mutaiton, and CNV which are prominent omics layer for analyzing and studying TCGA-BRCA. Third, visualizing more specific omics data for each patient. Adding to TF-TG correlations, scientists are able to get some ideas from integrated omics data on biological pathways.

In the future, to provide more helpful service for scientists, BioVLAB-TCGA could add some more visualization functions. Customer data availability are mostly needed. Since many bioinformatics tools are working with customer dataset, scientists could already have preprocessed various omics data other than TCGA's. With BioVLAB-TCGA, scientists can find more interesting result by navigating on KEGG pathways. In addition to this, other TCGA-dataset availability is also needed.

Bibliography

1. Aravindhan, G., Kumar, G. R., Kumar, R. S., and Subha, K. (2009). Ajax interface: a breakthrough in bioinformatics web applications. Proteomics Insights, 2, 1.
2. Azim, H. A., Nguyen, B., Broh e, S., Zoppoli, G., and Sotiriou, C. (2015). Genomic aberrations in young and elderly breast cancer patients. BMC medicine, 13(1), 1.
3. Benevolenskaya, E. V., Islam, A. B., Ahsan, H., Kibriya, M. G., Jasmine, F., Wolff, B., Al-Alem, U., Wiley, E., Kajdacsy-Balla, A., Macias, V., et al. (2016). Dna methylation and hormone receptor status in breast cancer. Clinical epigenetics, 8(1), 1.
4. Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., et al. (2010). Biocatalogue: a universal catalogue of web services for the life sciences. Nucleic acids research, page gkq394.
5. Bianco, L., Riccadonna, S., Lavezzo, E., Falda, M., Formentin, E., Cavalieri, D., Toppo, S., and Fontana, P. (2016). Pathway inspector: a pathway based web application for rnaseq analysis of model and non-model organisms. Bioinformatics, page btw636.
6. Bose, R., Kavuri, S. M., Searleman, A. C., Shen, W., Shen, D.,

Koboldt, D. C., Monsey, J., Goel, N., Aronson, A. B., Li, S., et al. (2013). Activating her2 mutations in her2 gene amplification negative breast cancer. Cancer discovery, 3(2), 224–237.

7. Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). Htridb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC genomics, 13(1), 405.
8. Creighton, C. J. et al. (2012). The molecular profile of luminal b breast cancer. Biologics, 6(2), 289–297.
9. Ellis, M. J. and Perou, C. M. (2013). The genomic landscape of breast cancer as a therapeutic roadmap. Cancer discovery, 3(1), 27–34.
10. Fielding, R. T. and Taylor, R. N. (2002). Principled design of the modern web architecture. ACM Transactions on Internet Technology (TOIT), 2(2), 115–150.
11. Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Science signaling, 6(269), pl1.
12. Gauthier, N. P., Reznik, E., Gao, J., Sumer, S. O., Schultz, N., Sander, C., and Miller, M. L. (2015). Mutationaligner: a resource of recurrent mutation hotspots in protein domains in cancer. Nucleic acids research, page gkv1132.
13. Grossman, R. L., Heath, A., Murphy, M., Patterson, M., and Wells, W. (2016). A case for data commons: Towards data science as a service. arXiv preprint arXiv:1604.02608.

14. Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27–30.
15. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The kegg resource for deciphering the genome. Nucleic acids research, 32(suppl 1), D277–D280.
16. Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe Ł TMer, I., LaFramboise, T., Brown, M., Tyekucheva, S., and Freedman, M. L. (2013). Integrative eqtl-based analyses reveal the biology of breast cancer risk loci. Cell, 152(3), 633–641. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsd ttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. Bioinformatics, 27(12), 1739–1740.
17. Luo, W. and Brouwer, C. (2013). Pathview: an r/bioconductor package for pathwaybased data integration and visualization. Bioinformatics, 29(14), 1830–1831.
18. Masse, M. (2011). REST API design rulebook. " O' Reilly Media, Inc.".
19. Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome biology, 12(4), 1.
20. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., et al. (2013). Large-scale genotyping

identifies 41 new loci associated with breast cancer risk. Nature genetics, 45(4), 353–361.

21. Moiseeva, E. P., Heukers, R., and Manson, M. M. (2006). Egfr and src are involved in indole-3-carbinol-induced death and cell cycle arrest of human breast cancer cells. Carcinogenesis, 28(2), 435–445.
22. Network, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumours. Nature, 490(7418), 61–70.
23. Network, N. C. C. et al.(2003). Breast cancer clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network: JNCCN, 1(2), 148.
24. Onitilo, A. A., Engel, J. M., Greenlee, R. T., and Mukesh, B. N. (2009). Breast cancer subtypes based on er/pr and her2 expression: comparison of clinicopathologic features and survival. Clinical medicine & research, 7(1-2), 4–13.
25. Prat, A., Adamo, B., Cheang, M. C., Anders, C. K., Carey, L. A., and Perou, C. M. (2013). Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. The oncologist, 18(2), 123–133.
26. Roberts, A. M., Wong, A. K., Fisk, I., and Troyanskaya, O. G. (2016). Giant api: an application programming interface for functional genomics. Nucleic acids research, page gkw289.
27. Robinson, J. T., Thorvaldsd ttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. Nature biotechnology, 29(1), 24–26.

28. Somerville, C. and Somerville, S. (1999). Plant functional genomics. Science, 285(5426), 380–383.
29. Stewart, B., Wild, C. P., et al. (2016). World cancer report 2014. World.
30. Trichopoulos, D., MacMahon, B., and Cole, P. (1972). Menopause and breast cancer risk. Journal of the National Cancer Institute, 48(3), 605–613.
31. Van De Vijver, M. J., He, Y. D., van' t Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine, 347(25), 1999–2009.
32. Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P. (2014). The ensembl rest api: ensembl data for any language. Bioinformatics, page btu613.

요약

BioVLAB-TCGA는 TCGA(The Cancer Genome Atlas)에 공개되어 있는 유방암 환자 1,098명으로부터 생성된 20,531개의 유전자 데이터를 KEGG 패스웨이에 매핑하여 색으로 시각화한 결과를 보여주는 플랫폼이다. 기존 생물정보연구자들은 TCGA 데이터를 분석하고 규명해야 할 유전자를 발견하는 등 연구를 진행할 때 생물학적인 전문지식에 의존해 왔다. 하지만 이번 연구를 통하여 생물학에서 검증된 결과를 통합하여 연구할 수 있는 환경을 구축한 것이다. KEGG 패스웨이를 이용해 TCGA에서 생물학적 의미를 가진 유전자군을 필터링하고, TCGA 임상 데이터에서 환자군을 필터링해 점차 구체적인 데이터에 접근 가능한 시스템을 개발했다.

특히 생물정보 분석에서 중요한 오믹스 데이터인 유전자 발현량, 복제수변이, 돌연변이에 대한 통합적 접근이 가능하다. BioVLAB-TCGA를 통해 오믹스 데이터를 KEGG 패스웨이 위에 시각화 할 수 있고, 이 세 데이터를 한 유전자 엔트리에 표시해 오믹스 데이터의 특정 패턴을 찾아낼 수 있다.

BioVLAB-TCGA에서는 TCGA 데이터를 모델링하여 만든 데이터베이스와 KEGG REST에서 얻은 패스웨이 이미지를 연동한다. TCGA 유방암 환자의 데이터를 임상정보를 토대로 필터링하고, 그 결과를 손쉽게 KEGG 패스웨이에 표현해 생물 전문가들이 향후 연구를 위한 유전자 후보군을 발굴할 수 있다.