BMC Bioinformatics

METHODOLOGY Open Access

BRCA-Pathway: a structural integration and visualization system of TCGA breast cancer data on KEGG pathways

Inyoung Kim[1], Saemi Choi[1] and Sun Kim[1,2,3*]

From Proceedings of the 28th International Conference on Genome Informatics: bioinformatics
Seoul, Korea. 31 October - 3 November 2017

Abstract

Background: Bioinformatics research for finding biological mechanisms can be done by analysis of transcriptome data with pathway based interpretation. Therefore, researchers have tried to develop tools to analyze transcriptome data with pathway based interpretation. Over the years, the amount of omics data has become huge, e.g., TCGA, and the data types to be analyzed have come in many varieties, including mutations, copy number variations, and transcriptome. We also need to consider a complex relationship with regulators of genes, particularly Transcription Factors(TF). However, there has not been a system for pathway based exploration and analysis of TCGA multi-omics data. In this reason, We have developed a web based system BRCA-Pathway to fulfill the need for pathway based analysis of TCGA multi-omics data.

Results: BRCA-Pathway is a structured integration and visual exploration system of TCGA breast cancer data on KEGG pathways. For data integration, a relational database is designed and used to integrate multi-omics data of TCGA-BRCA, KEGG pathway data, Hallmark gene sets, transcription factors, driver genes, and PAM50 subtypes. For data exploration, multi-omics data such as SNV, CNV and gene expression can be visualized simultaneously in KEGG pathway maps, together with transcription factors-target genes (TF-TG) correlation and relationships among cancer driver genes. In addition, 'Pathways summary' and 'Oncoprint' with mutual exclusivity sort can be generated dynamically with a request by the user. Data in BRCA-Pathway can be downloaded by REST API for further analysis.

Conclusions: BRCA-Pathway helps researchers navigate omics data towards potentially important genes, regulators, and discover complex patterns involving mutations, CNV, and gene expression data of various patient groups in the biological pathway context. In addition, mutually exclusive genomic alteration patterns in a specific pathway can be generated. BRCA-Pathway can provide an integrative perspective on the breast cancer omics data, which can help researchers discover new insights on the biological mechanisms of breast cancer.

Keywords: TCGA breast cancer, KEGG-Pathway, Gene expression, Mutation, Copy number variation

Background

Transcriptome data measured at the whole genome level requires interpretation at a higher level. For this reason, biological pathway analysis of transcriptome data has become a standard approach, e.g., Pathview [1], Pathway Inspector [2]. However, existing pathway based analysis tools are not powerful enough for the analysis of omics data from cancer, a complex disease that requires integrated analysis of multi-omics data. For example, single nucleotide variation (SNV) and copy number variation (CNV) are frequently measured for cancer research. Thus, it is necessary to integrate and analyze multi-omics datasets of different types together. A number of tools, including TCGA2STAT [3], have been developed for TCGA multi-omics analysis. However, these tools require experienced programming abilities and knowledge about the detailed specification of TCGA data. Meanwhile, there

*Correspondence: sunkim.bioinfo@snu.ac.kr
[1] Department of Computer Science and Engineering, Seoul National University, 1 Gwanak-ro, 08826 Seoul, Korea
[2] Interdisciplinary Program in Bioinformatics, Seoul National University, 1 Gwanak-ro, 08826 Seoul, Korea
Full list of author information is available at the end of the article

exist web-based multi-omics data analysis services, such as cBioPortal [4], OASIS [5], NetGestalt [6]. These frameworks provide easy access and interpretation of multi-omics data. OASIS provides multi-omics data such as mutation, CNV, and gene expression in selected cancer types and selected oncogenic pathways. However, multi-omics data is presented in a table format, so there is no information about the relationship between genes. NetGestalt provides network-centric view of multi-omics data by adopting visualization on the horizontal dimension to scale up to large networks. In addition, by zooming into a specific gene, multi-omics data of each gene is shown and protein-protein interactions around the selected gene are provided. KeyPathwayMinerWeb provides online multi-omics network enrichment analysis. From user provided gene expression data and an active gene list, maximally connected subnetworks using PPI are provided. These systems display the relationship of genes on the network but do not present multi-omics data on the network. On the contrary, our system, BRCA-Pathway, is designed to represent relationship between genes and corresponding multi-omics data simultaneously on the KEGG pathways. In summary, pathway based multi-omics analysis system is necessary but challenging due to larger sample sizes and higher dimension (multi-omics). For these reasons, we developed BRCA-Pathway, a web-based interactive exploration and visualization system of TCGA breast cancer data on KEGG pathways to provide broad perspective of TCGA breast cancer data. The major features are:

- Multi-omics data such as SNV, CNV, and gene expression can be visualized simultaneously on KEGG pathway maps, together with TF-TG correlation and relationships among cancer driver genes.
- Users can perform comparative analysis of BRCA data, including selection of differentially expressed genes (DEGs) in arbitrary patient groups, mutual exclusivity module (MEMo) summary of genomic alterations (SNV and CNV).
- Data can be downloaded by REpresentational State Transfer Application Programming Interface (REST API).

Results

BRCA-Pathway system design

BRCA-Pathway consists of three components: Database system, REST API, and Web front-end. Overall system design is described in Fig. 1. BRCA-Pathway integrates multiple resources in a relational database and provides a web-based interactive interface and REST API. A database system using MySQL is designed for the structural integration of multi-omics data of TCGA-BRCA, KEGG pathway data, Hallmark gene sets, TF-TG relationships, driver genes, and PAM50 subtype. BRCA-Pathway system can update KEGG Pathways and TCGA-BRCA data by initiating the update software module to incorporate the most recent information. Current configuration is based on KEGG Pathway released on October 1, 2016 and the latest version of GDAC Firehose on January 28, 2016.

Data sources

BRCA-Pathway stores and utilizes data such as TCGA breast cancer data, breast cancer subtype by PAM50, TF-TG regulation data, Hallmark gene sets, driver gene list and KEGG pathway data. Data is accessible by web-server system or REST API. Data source and status is described in the Table 1.

TCGA breast cancer multi-omics data

TCGA breast cancer data was obtained from FIREHOSE [7]. Clinical data was obtained from 'Clinical_Pick_Tier1'. In addition to clinical data common to all cancer types, breast cancer specific features were selected and stored in BRCA-Pathway system. Additional features are ER, PR, HER2 receptor status and menopause status. Three hormone receptors have been used as the marker of the breast cancer sub-typing [8]. Also, late menopause at age over 55 is known as an risk factor of breast cancer [9]. Gene expression data was obtained from 'illuminahiseq_rnaseqv2-RSEM_genes_normalized'. Each breast cancer patient has 20,531 gene-level RNA-seq expression data (RSEM normalized counts) from their tumor samples. About 10% of the patient have gene expression data from their normal samples. We used gene expression data from patient wide normal samples as the normal pool data, and gene expression data from patient wide tumor samples as the tumor pool data for the comparison and visualization of gene expression data from selected sub-population. Mutation data was obtained from 'Mutation_Packager_Oncotated_Calls'. The mutation table includes mutated gene ID, and detailed information about the mutation. Because the patterns of mutation in oncogenes are nonrandom and characteristic, and also because oncogenes are recurrently mutated at the same amino acid positions [10], we included detailed information such as 'genome_change' field describing the chromosomal position. CNV data was obtained from 'CopyNumber Gistic2'.

Breast cancer subtype data

Breast cancers are classified into subtypes based on gene expression data. PAM50 breast cancer subtyping is widely used method to classify breast cancer into four subtype: Luminal A, Luminal B, Basal-like, HER2-enriched. Breast cancer subtype data is generated using PAM50 predictor bioclassifier R script [11]. The gene expression value of each of PAM50 genes was log2-transformed before the

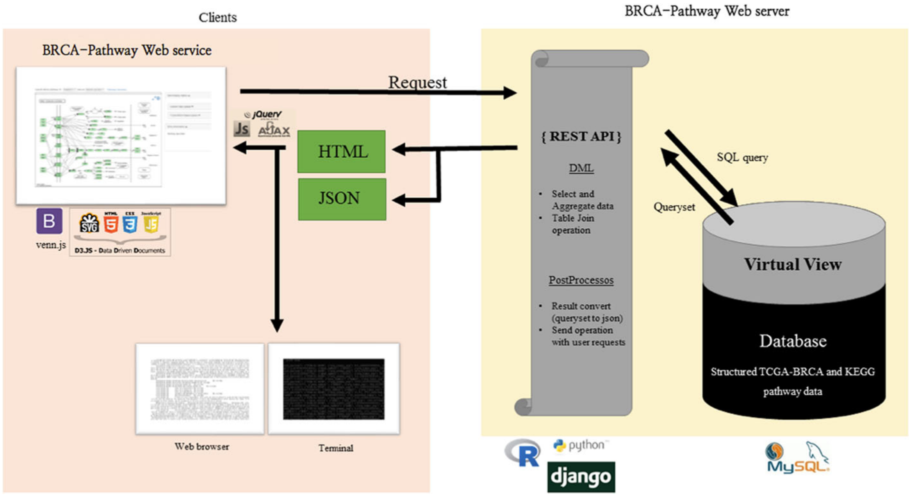


Fig. 1 System design for BRCA-Pathway. BRCA-Pathway system structure overview. The left part is client side, and the right part is BRCA-Pathway server side. BRCA-Pathway server builds the database which contains structured TCGA multi-omics data and KEGG pathway data. Database is abstracted by virtual view which simplifies table join functions. REST API could get the data set when it throws query to database. These data set will be provided by Django framework so that the clients can access the data set by web browser or terminal. BRCA-Pathway web front-end receives the data set and visualizes for clients

PAM50 run. Users can select the patient group by subtype and TCGA clinical information.

KEGG pathway data

BRCA-Pathway stored KEGG pathway data by accessing KEGG API [12] and KEGG PATHWAY Database [13]. Gene information such as gene names, KEGG IDs, and relations between pathways and genes was obtained from KEGG REST API [14]. Information about graphical representations of pathway was acquired from KEGG PATHWAY Database [15].

TF-TG regulation data

KEGG pathway includes information about the gene-gene relation (activation, suppression, gene-gene interaction, etc.). However, relationships between TFs and potential target genes are not provided by the KEGG pathway. To supplement the KEGG pathway, we provide Pearson correlation coefficient between TF-TG with the KEGG pathway. TF-TG relationships were obtained from two different databases. Molecular Signatures Database provides gene sets that share a transcription factor binding site defined in TRANSFAC database [16]. Human Transcriptional Regulation Interaction Database (HTRIdb) is a database for experimentally verified human transcriptional regulation interactions [17]. The union set of both databases are stored in BRCA-Pathway TF-TG table.

Hallmark gene sets

Hallmark gene sets summarize and represent well-defined biological states or processes and display coherent expression. Fifty Hallmark gene sets were obtained from MSigDB [18]. When researchers are interested in specific biological process, they can start to explore TCGA data by selecting Hallmark gene sets.

Driver gene list

Driver gene is a gene that contains driver gene mutations or is expressed aberrantly in a fashion that confers a selective growth advantage [10]. We obtained driver gene list consisting of oncogenes and Tumor Suppressor Genes (TSG) from Cancer Gene Census in COSMIC database [19] and the previous research [10]. BRCA-Pathway represents driver genes on the pathway.

BRCA-Pathway workflow

In order to start with BRCA-Pathway, users need to select pathways of their interest. There are several ways to select pathways (Fig. 2). First, users select pathways by differentially expressed genes (DEGs). Users select patient sub-population by PAM50 subtype and clinical information, then BRCA-Pathway starts to perform a Wilcoxon rank-sum test for two groups, i.e., selected sub-population as case and 112 normal samples or remaining rest tumor samples as control. As a result, DEGs and KEGG pathways including DEGs are listed. For example, when users want to explore 'Basal-like' subtype patients and select the condition for DEG computation, BRCA-Pathway returns the pathway list including DEGs in 'Basal-like' subtype compared with either normal samples or the rest breast tumor samples. After selecting pathways and loading TCGA data, BRCA-

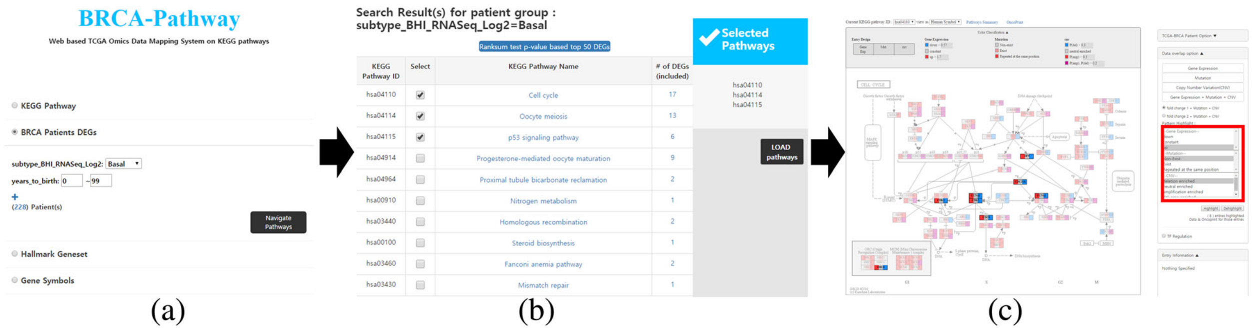
Table 1 Description of data provided by BRCA-Pathway: data type, data size, and data source

Data type	Data size	Data source
Clinical	1098 patients 22 attributes 1098 rows	Standardized analysis-ready TCGA data, Broad Institute TCGA genome
Expression	1093 patients 20,531 genes 22,440,383 rows	data analysis center (2016): Firehose stddata__2016_01_28 run. Broad Institute of MIT and
Mutation	977 patients 17,280 genes 86,765 rows	Harvard. doi:10.7908/C11G0KM9
CNV	1080 patients 24,776 genes 26,758,080 rows	Broad Institute TCGA genome data analysis center (2016): SNP6 copy number analysis (GISTIC2). Broad Institute of MIT and Harvard. doi:10.7908/C1NP23RQ
KEGG pathway	307 pathways	Kyoto encyclopedia of genes and genomes (KEGG)
Transcription factors	85,314 TF-TG pairs	Human transcriptional regulation interaction database (HTRIdb) Molecular signatures database (MSigDB)
Hallmark gene sets	50 Hallmark gene sets	Molecular signatures database (MSigDB)
Driver genes	486 driver genes	Cancer gene census in COSMIC database Vogelstein, Bert, et al. “Cancer genome landscapes.” science 339.6127 (2013): 1546-1558. Table S2A

Pathway will show gene-expression, CNV and mutation data mapped on selected pathways. Second, users select pathways from Hallmark gene sets. When a hallmark gene set is selected, BRCA-Pathway shows the list of related KEGG pathways by the number of included genes in the selected hallmark gene set. In the same way, users are provided KEGG pathway list by submitting gene symbols.

Pathway-based exploration of TCGA BRCA

BRCA data can be explored in different ways. For example, when users specify subpopulation of TCGA breast cancer patients, the system loads multi-omics data of the selected subpopulation into the web-browser. By clicking ‘Data overlap option’, users can change the type of data (gene expression, mutation and CNV) mapped onto the pathway. This enables users to see the same pathway by three different points of view. BRCA-Pathway colorizes KEGG Pathway entries according to the user-controllable classification criteria. The system provides a function that highlights specific patterns of omics data such as gene expression level up, mutation free and copy number deletion (Fig. 3). In addition, ‘Pathways summary’ and ‘Oncoprint’ are available. Pathways summary is shown in Venn diagram [20] that shows overlapped genes among selected multiple pathways (Fig. 4). Genomic alterations such as mutation and CNV in the same pathway are often mutually exclusive [21] and different combination of alterations are sufficient to perturb the pathways [22]. Thus, predicting driver alterations from the frequency of occurrence is not easy. With pathway leveled view, overall trend of genomic alteration is shown, but mutually exclusive patterns of alteration are not recognizable. To compensate this problem, we provide OncoPrint [23] with ‘mutual

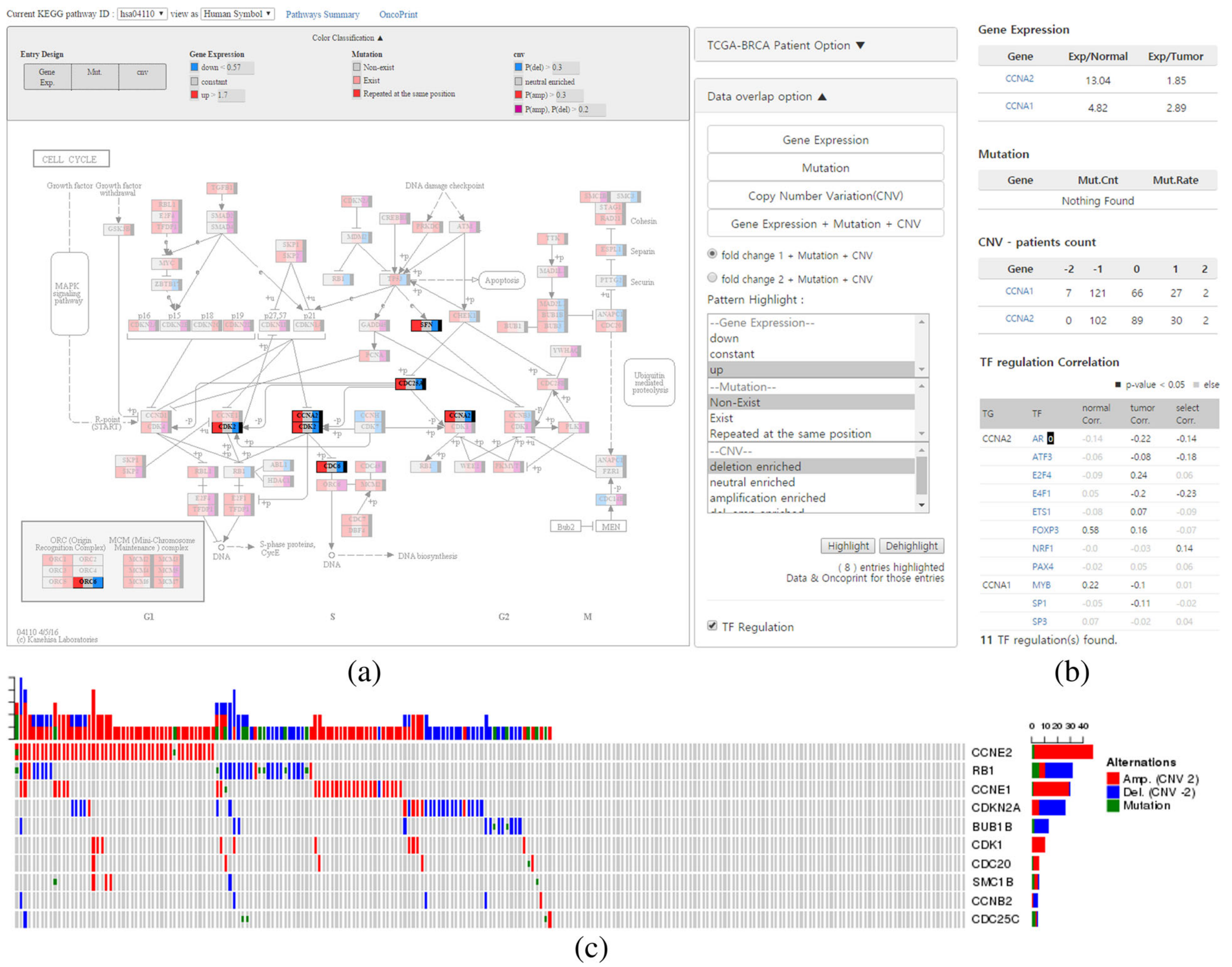


Search Result(s) for patient group : subtype_BHI_RNASeq_Log2=Basal

Ranksum test p-value based top 50 DEGs

KEGG Pathway ID	Select	KEGG Pathway Name	# of DEGs (included)
hsa04110	☑	Cell cycle	17
hsa04114	☑	Oocyte meiosis	13
hsa04115	☑	p53 signaling pathway	6
hsa04914	□	Progesterone-mediated oocyte maturation	9
hsa04964	□	Proximal tubule bicarbonate reclamation	2
hsa00910	□	Nitrogen metabolism	1
hsa03440	□	Homologous recombination	2
hsa00100	□	Steroid biosynthesis	1
hsa03460	□	Fanconi anemia pathway	2
hsa03430	□	Mismatch repair	1

Fig. 2 Example workflow of BRCA-Pathway. There are several ways to select KEGG pathways. In this example, pathway selection by DEG is described. First, select subpopulation by breast cancer subtype and clinical features (**a**). If patients with ‘Basal-like’ subtype are selected, then pathways including differentially expressed genes are listed as the result of computation of Wilcoxon rank-sum test (**b**). Select pathways from the list then BRCA-Pathway visualizes multi-omics data of patients whose subtype is ‘Basal-like’. Users can view the pathway by three different data types. Gene expression, mutation, and CNV data are visualized in separate way or integrated way. Genes having specific omics-pattern are highlighted by selecting the omics condition. For example, genes having no mutation and of which CNV is loss or del but overexpressed are selected for highlighting then the pathway turns shaded except the highlighted genes (**c**)



Current KEGG pathway ID : hsa04110 ▾ view as Human Symbol ▾ Pathways Summary OncoPrint

Color Classification ▲

Entry Design: Gene Exp. | Mut. | cnv

Gene Expression: 🟦 down < 0.57; □ constant; 🟥 up > 1.7

Mutation: □ Non-exist; 🟥 Exist; 🟥 Repeated at the same position

cnv: 🟦 P(del) > 0.3; □ neutral enriched; 🟥 P(amp) > 0.3; 🟪 P(amp), P(del) > 0.2

TCGA-BRCA Patient Option ▼

Data overlap option ▲

Gene Expression

Mutation

Copy Number Variation(CNV)

Gene Expression + Mutation + CNV

● fold change 1 + Mutation + CNV

○ fold change 2 + Mutation + CNV

Pattern Highlight :

--Gene Expression--
down
constant
up

--Mutation--
Non-Exist
Exist
Repeated at the same position

--CNV--
deletion enriched
neutral enriched
amplification enriched

Highlight Dehighlight

(8) entries highlighted
Data & Oncoprint for those entries

☑ TF Regulation

Gene Expression

Gene	Exp/Normal	Exp/Tumor
CCNA2	13.04	1.85
CCNA1	4.82	2.89

Mutation

Gene	Mut.Cnt	Mut.Rate
Nothing Found		

CNV - patients count

Gene	-2	-1	0	1	2
CCNA1	7	121	66	27	2
CCNA2	0	102	89	30	2

TF regulation Correlation

■ p-value < 0.05 ■ else

TG	TF	normal Corr.	tumor Corr.	select Corr.
CCNA2	AR 0	-0.14	-0.22	-0.14
	ATF3	-0.06	-0.08	-0.18
	E2F4	-0.09	0.24	0.06
	E4F1	0.05	-0.2	-0.23
	ETS1	-0.08	0.07	-0.09
	FOXP3	0.58	0.16	-0.07
	NRF1	-0.0	-0.03	0.14
	PAX4	-0.02	0.05	0.06
CCNA1	MYB	0.22	-0.1	0.01
	SP1	-0.05	-0.11	-0.02
	SP3	0.07	-0.02	0.04

11 TF regulation(s) found.

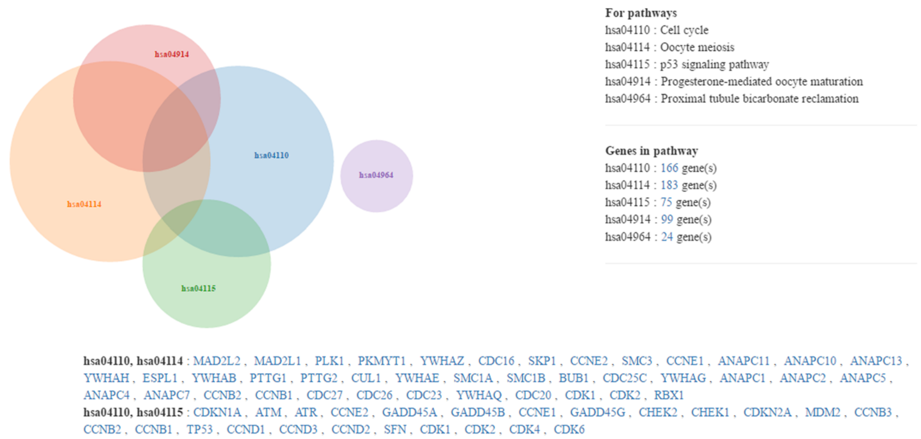
Fig. 3 TCGA BRCA multi-omics data exploration with BRCA-Pathway. Visualization of TCGA BRCA with BRCA-Pathway (**a**) Multi-omics data of patients whose subtypes are 'Basal-like' is mapped on hsa04110 (Cell cycle). Users can view pathway maps in three different data types. Gene expression, mutation and CNV data are visualized either separately or integrated. Genes with specific patterns are highlighted by selecting the conditions. For example, genes, having no mutation and of which CNV is loss or del but that are over-expressed, are selected. Genes in the pathway except the highlighted genes become shaded (**b**) Clicking a gene entry on the pathway displays multi-omics data of the gene such as gene expression level compared with normal sample pool and with tumor sample pool, mutation count and ratio, CNV as -2 (del), -1 (loss), 0 (neutral), 1 (gain), 2 (amp) with the count of patients having each CNV value and Pearson correlation coefficients of TF-TG gene expression. Three columns represent the correlation of TF-TG in normal sample pool, tumor sample pool and selected subpopulations pool, respectively. Correlation with *p*-value over 0.05 are shaded. **c** Selecting genes, Oncoprint shows copy number variation (amp or del) and mutation in the mutual exclusive way between selected genes. In this case, some extremely over-expressed genes (fold change>10 by normal sample pool) and RB1 are selected and the figure shows mutual exclusive pattern between CCNE2, RB1, CCNE1 and CDKN2A (top 4 rows respectively)

exclusivity sort'. Users can check on mutual exclusivity in selected patient group with respect to CNV and mutation.

Figure 3a shows a pathway leveled view of 'Basal-like' subtype patients and Fig. 3b shows the individual omics data of CCNA1, CCNA2. The highlighted genes are mutation free, copy number loss or deletion enriched but over-expressed compared to normal pool data. Because mutation effects are excluded, there probably exist regulators that promoted the transcription of CCNA1, CCNA2. With provided TF-TG correlation coefficient, users get TF list targeting CCNA1, CCNA2 and see that the correlation coefficients are different between normal samples and selected tumor samples.

User data visualization on KEGG pathways

BRCA-Pathway provides visualization of user data to extend the usability of the system. After switching to 'User data mode', input a text file consisting of Entrez geneId and fold change value, and the gene expression level is shown in color. Adjusting color or threshold value helps to customize pathway visualization. However, unlike 'TCGA



For pathways
hsa04110 : Cell cycle
hsa04114 : Oocyte meiosis
hsa04115 : p53 signaling pathway
hsa04914 : Progesterone-mediated oocyte maturation
hsa04964 : Proximal tubule bicarbonate reclamation

Genes in pathway
hsa04110 : 166 gene(s)
hsa04114 : 183 gene(s)
hsa04115 : 75 gene(s)
hsa04914 : 99 gene(s)
hsa04964 : 24 gene(s)

hsa04110, hsa04114 : MAD2L2 , MAD2L1 , PLK1 , PKMYT1 , YWHAZ , CDC16 , SKP1 , CCNE2 , SMC3 , CCNE1 , ANAPC11 , ANAPC10 , ANAPC13 , YWHAH , ESPL1 , YWHAB , PTTG1 , PTTG2 , CUL1 , YWHAE , SMC1A , SMC1B , BUB1 , CDC25C , YWHAG , ANAPC1 , ANAPC2 , ANAPC5 , ANAPC4 , ANAPC7 , CCNB2 , CCNB1 , CDC27 , CDC26 , CDC23 , YWHAQ , CDC20 , CDK1 , CDK2 , RBX1
hsa04110, hsa04115 : CDKN1A , ATM , ATR , CCNE2 , GADD45A , GADD45B , CCNE1 , GADD45G , CHEK2 , CHEK1 , CDKN2A , MDM2 , CCNB3 , CCNB2 , CCNB1 , TP53 , CCND1 , CCND3 , CCND2 , SFN , CDK1 , CDK2 , CDK4 , CDK6

Fig. 4 Venn diagram for Pathways summary. Pathway relations are shown in Venn diagram. A list of pathways and the number of genes in each pathway are described on the right side. Genes shared between pathways are listed under the Venn diagram

mode', only a single gene selected by the entry label is considered and all the other genes belonging to the entry are ignored.

REST API

REST API separates data extraction from the developmental environment so that users can easily extract data without understanding the internal system [24]. Given patient selection option in the web page, a front-end program creates a URL and sends the URL to REST API. This allows the system to aggregate omics data sets for a subset of patients to create a dynamic user view. By using REST API, it is possible to extract genes contained in KEGG pathway maps and to aggregate TCGA data after receiving query result from MySQL. Users can access to data with simple endpoint coding.

The domain address is the server URL that BRCA-Pathway is configured on. After the slash(/) mark, at least one argument should be given. The 1st argument specifies the data to retrieve and the argument can be `landscape`, `search`, `genes`, `pathways`, `TCGA-BRCA`. In case of `tcga-brca.bhi2.snu.ac.kr/api/landscape`, 'landscape' represents the current status of TCGA data and KEGG pathway. 'search' means that the pathway list will be provided by searching gene names or pathway names, and 'genes' provides the gene list in pathways specified by argument 2. Furthermore, 'pathways' returns pathway information specified by argument 3's endpoint filtered by argument 2. REST API examples are listed below. The last example means that 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.

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

Patient_options are listed below:

- subtype_BHI_RNASeq_Log2 : all|*Basal*|Her2|*LumA*|*LumB*|Normal
- years_to_birth_from : integer & years_to_birth_to : integer
- er_status : all|*indeterminate*|*negative*|positive
- pr_status : all|*indeterminate*|*negative*|positive
- her2_status : all|*indeterminate*|*negative*|*positive*|equivocal
- vital_status : all|0|1 *0: alive, 1:dead
- pathologic_stage : all|stage_i|stage_ii|stage_iii|stage_iv|stage_iv|stage_tis|stage_x
- pathologic_T_stage : all|t1|t2|t3|t4|tx
- pathologic_N_stage : all|n0|n1|n2|n3|nx
- pathologic_M_stage : all|cm0_|m0|m1|mx
- gender : all|*female*|male
- radiation_therapy : all|*no*|yes
- histological_type : all|infiltrating_carcinoma_nos|infiltrating_ductal_carcinoma|infiltrating_lobular_carcinoma|medullary_carcinoma|metaplastic_carcinoma|mixed_history (please_specify)|mucinous_carcinoma|other__specify
- number_of_lymph_nodes : all|0|1|2 *1: #of node less than or equal to 10, 2: greater thane 10

Table 2 REST API arguments and parameters supported by BRCA-Pathway

arg1	arg2	arg3	Parameters	Description
landscape	-	-	type=[JSON | LIST] fields=[KEGGPathway_update | KEGGPathways | Pathway_Pathway_Membership | TCGAGenes | TCGA_BRCA_Patients | TCGA_BRCA_update | TFTG_relation | driver_genes]	KEGG pathways, TCGA-BRCA data status
search	-	-	keyword type=[JSON | LIST] fields=[keyword | gene_labels | rate | pathID | pathname]	Get search result by keyword. Gene or pathway names with input_type.
genes	(pathIDs, sep="+")	- count	type=[JSON | LIST] fields=[pathID | gene_symbol | pathname | geneID | count**] **given only arg3 related_pathways	Retrieve pathways information. related_pathways operation supported.
pathways	(pathIDs, sep="+")	- related_pathways	type=[JSON | LIST] fields=[pathID | pathname | category_level2_id | description | related_pathID**] **given only arg3 related_pathways	Retrieve pathways information. related_pathways operation supported.
TCGA-BRCA	patient_list patient_count (pathIDs, sep="+")	- - (given pathIDs,) fold_change mutation CNV	type=[JSON | LIST] ** arg2 patient_list fields=[patient_barcode] ** arg2 patient_count fields=[count] ** arg3 fold_change fields=[geneID|gene_symbol | exp_select | exp_tumor | exp_normal | fc_select_tumor| fc_select_normal] ** arg3 mutation fields=[geneID | gene_symbol |count]	Retrieve pathways and TCGA-BRCA omics data. Gene expression with fold_change value, mutation with the number of patients having mutation, and CNV with GISTIC2 result range from -2 to 2. Each omics result will be given by gene level aggregating patients group omics data patient_list, patient_count operation supported.

Table 2 REST API arguments and parameters supported by BRCA-Pathway (*Continued*)

arg1	arg2	arg3	Parameters	Description
			** arg3 CNV fields=[geneID | gene_symbol | cnv_pos | cnv_zero | cnv_neg] ** And patient_options described_above	

- race : all|american_indian_or_alaska_native|*asian*|black_or_african_american|white
- ethnicity : all|hispanic_or_latino|not_hispanic_or_latino
- menopause_status : all|*indeterminate*|*peri*|*post*|pre

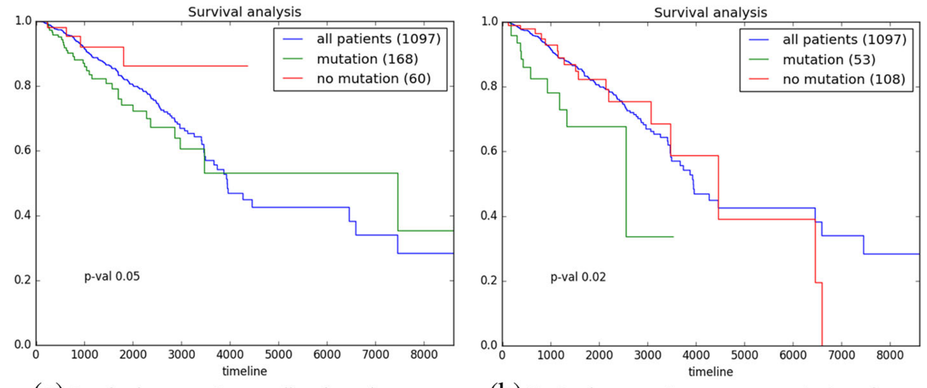
Discussions

Survival analysis

BRCA-Pathway provides survival analysis of selected patients. It divides the selected patients into two groups according to the presence of mutation in genes belong to a particular pathway and provides survival analysis for the patient population. For example, if user selected 'Basal' subtype and 'Cell cycle' pathway then patients of 'Basal' subtype are divided into two groups, mutation group and mutation free group. Patients having at least one mutation in the genes involved in the 'Cell cycle' pathway belong to the mutation group. On the other hand, patients without a mutation in 'Cell cycle' pathway belong to the mutation free group. Two hundred twenty-eight patients of 'Basal' subtype are divided into mutation group (168 patients) and mutation free group (60 patients). The green and red line represent the survival curves of mutation group and mutation free group, respectively. *P*-value by the logrank test is provided for the comparison of two groups, and all breast cancer patients are depicted as blue line for the convenience. Figure 5a gives a *p*-value of 0.05, which indicates a significant difference between the survival curves. Figure 5b gives a *p*-value of 0.02, the survival curve of 'Her2' patients having mutation in 'Oocyte meiosis' and 'Her2' patients having no mutation in 'Oocyte meiosis' pathway.

Gene prioritization

If the user wants to prioritize genes that is significant in 'Basal' subtype breast cancer, the exploration can start from DEGs in 'Basal' subtype. By selecting patient subpopulation as 'Basal' and comparison condition as normal pool, then DEGs and related pathways are shown. Since the 'Cell cycle' pathway is listed at the top rank, it is natural to select and load 'Cell cycle' pathway for the next step.



(a) Basal subtype patients, cell cycle pathway (b) Her2 subtype patients, oocyte meiosis pathway

Fig. 5 Survival analysis. **a** 'Basal' subtype and 'Cell cycle' pathway are selected, patients with at least one mutation in the genes involved in the 'Cell cycle pathway will belong to the mutation group (green line). On the other hand, patients without a mutation in 'Cell cycle' genes will belong to the mutation free group (red line). **b** 'Her2' subtype and 'Oocyte meiosis' pathway are selected, 'Her2' patients having at least one mutation in 'Oocyte meiosis' genes are depicted as green line, and the rest 'Her2' patients are depicted as red line. *P*-value by the logrank test is provided for the comparison of two groups, and all breast cancer patients are depicted as blue line for the convenience

After loading a pathway, the user can explore multi-omics data on the 'Cell cycle' pathway of 'Basal' subtype patients. Figure 3a shows a pathway level view of gene expression, mutation and copy number variation. By highlight function, user can locate genes with the specified condition. The highlighted genes are mutation free, copy number loss or deletion enriched but over-expressed compared to normal pool data. Individual multi-omics data is obtained by clicking each entry and Fig. 3b shows the multi-omics data of CCNA1, CCNA2. Because mutation effects are excluded, it would be interesting to see if there exist regulators that promoted the transcription of CCNA1 and CCNA2. With 'TF regulation' checked, TF-TG correlation coefficients are provided and then user can get the TFs targeting CCNA1, CCNA2 and see that the correlation coefficient between FOXP3 and CCNA2 is much different in normal samples (0.58) and tumor samples (0.16). Since p-value is over 0.05, correlation coefficient in 'Basal' subtype samples (-0.07) is shaded. Although we could not identify the responsible genes that promoted transcription of CCNA1 and CCNA2, we found that positive correlation between FOXP3 as TF and CCNA2 as TG is disrupted in tumor samples.

Conclusions

BRCA-Pathway helps researchers navigate multi-omics data towards potentially important genes, regulators, and discover complex patterns involving mutations, CNV, and gene expression data of various patient groups in the biological pathway context. In addition, mutually exclusive genomic alteration patterns in a specific pathway can be generated. BRCA-Pathway can provide an integrative perspective on the breast cancer omics data, which can help researchers discover new insights on the biological mechanisms of breast cancer. In the future, BRCA-Pathway could include other omics data sets such as miRNA expression and DNA promoter methylation profiles to support more extensive research. And besides breast cancer data of TCGA, other cancer dataset availability is also needed.

Abbreviations

CNV: Copy number variation; DEG: Differentially expressed genes; HTRIdb: Human transcriptional regulation interaction database; KEGG: Kyoto encyclopedia of genes and genomes; MEMo: Mutual exclusivity module; MSigDB: Molecular signatures database; PAM50: Prediction analysis of microarray 50; REST-API: Representational state transfer application programming interface; RNA: Ribonucleic acid; RNA-seq: Whole transcriptome sequencing; SNV: Single nucleotide variation; TCGA: The cancer genome atlas; TF: Transcription factor; TG: Target gene; TSG: Tumor suppressor gene

Acknowledgements

Authors are grateful to anonymous reviewers for their useful comments on the manuscript.

Funding

This work was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI15C3224), and Collaborative Genome Program for Fostering New Post-Genome industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science ICT and Future Planning (No.NRF-2014M3C9A3063541), and Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.B0717-16-0098, Development of homomorphic encryption for DNA analysis and biometry authentication). The publication cost was funded by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.B0717-16-0098, Development of homomorphic encryption for DNA analysis and biometry authentication).

Availability of data and materials

The datasets are freely available and accessible at http://biohealth.snu.ac.kr/software/brcapathway. Data used for the study can be downloaded from BROAD FireBrowse (Refer to Table 1).

About this supplement

This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 1, 2018: Proceedings of the 28th International Conference on Genome Informatics: bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-1.

Authors' contributions

SK designed and supervised the research project and edited the paper. IY designed database system and managed the development of the system and wrote the manuscript. SM developed web service including REST API and helped to write the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science and Engineering, Seoul National University, 1 Gwanak-ro, 08826 Seoul, Korea. [2]Interdisciplinary Program in Bioinformatics, Seoul National University, 1 Gwanak-ro, 08826 Seoul, Korea. [3]Bioinformatics Institute, Seoul National University, 1 Gwanak-ro, 08826 Seoul, Korea.

Published: 19 February 2018

References

1. Luo W, Brouwer C. Pathview: an r/bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;29(14):1830–1.
2. Bianco L, Riccadonna S, Lavezzo E, Falda M, Formentin E, Cavalieri D, Toppo S, Fontana P. Pathway inspector: a pathway based web application for rnaseq analysis of model and non-model organisms. Bioinformatics. 2016;33(3):453–455.
3. Wan YW, Allen GI, Liu Z. Tcga2stat: simple tcga data access for integrated statistical analysis in R. Bioinformatics. 2015;32(6):952–954.
4. cBioPortal. http://www.cbioportal.org/. Accessed 13 Feb 2017.
5. Fernandez-Banet J, Esposito A, Coffin S, Horvath IB, Estrella H, Schefzick S, Deng S, Wang K, AChing K, Ding Y, et al. Oasis: web-based platform for exploring cancer multi-omics data. Nat Methods. 2016;13(1):9.
6. Zhu J, Shi Z, Wang J, Zhang B. Empowering biologists with multi-omics data: colorectal cancer as a paradigm. Bioinformatics. 2014;31(9):1436–43.
7. FIREHOSE Broad GDAC. http://gdac.broadinstitute.org/. Accessed 13 Feb 2017.
8. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. Breast cancer subtypes based on er/pr and her2 expression: comparison of clinicopathologic features and survival. Clin Med Res. 2009;7(1-2):4–13.

9. Trichopoulos D, MacMahon B, Cole P. Menopause and breast cancer risk. J Natl Cancer Inst. 1972;48(3):605–13.
10. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58.
11. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7.
12. KEGG API. http://www.kegg.jp/kegg/docs/keggapi.html. Accessed 7 Aug 2017.
13. KEGG Pathway. http://www.genome.jp/kegg/pathway.html. Accessed 7 Aug 2017.
14. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The kegg resource for deciphering the genome. Nucleic Acids Res. 2004;32(suppl 1):277–80.
15. Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
16. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
17. Bovolenta LA, Acencio ML, Lemke N. Htridb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC Genomics. 2012;13(1):405.
18. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
19. Cancer Gene Census. https://cancer.sanger.ac.uk/census/. Accessed 23 Aug 2017.
20. VENN Diagram. https://github.com/benfred/venn.js/. Accessed 7 Aug 2017.
21. Leiserson MD, Wu HT, Vandin F, Raphael BJ. Comet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. Genome Biol. 2015;16(1):1.
22. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22(2):398–406.
23. ComplexHeatmap. https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html. Accessed 13 Feb 2017.
24. Fielding RT, Taylor RN. Principled design of the modern web architecture. ACM Trans Internet Technol (TOIT). 2002;2(2):115–50.