LinkedOmics

A Web-based platform for analyzing cancer-associated multi-dimensional data

Manual

First edition 4 April 2017
Updated on July 3, 2017

LinkedOmics is a publicly available portal (http://linkedomics.org/) that includes multi-omics data from 32 TCGA cancer types. The platform includes data from methylation (gene level), copy number variation (focal and gene level), mutation (site and gene level), mRNA expression (gene level), miRNA expression (gene level), RPPA (analyte and gene level) and clinical data (phenotype) related to primary tumors from 11,158 patients. It also includes mass spectrometry-based proteomics data generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) for TCGA breast, colorectal, and ovarian tumors.

Note:
LinkedOmics is best run with the browser Chrome Version 56.0.2924.87+. If you are using Safari, then enable the hidden Develop menu: Pull down the "Safari" menu and choose "Preferences". Click on the "Advanced" tab. Check the box next to "Show Develop menu in menu bar". Click Develop menu on top panel and select Chrome/Firefox as the browser

Contents

1. Introduction

1.1 Overview

LinkedOmics provides a unique platform for biologists and clinicians to access, analyze, and compare cancer multi-omics data within and across tumor types. The web application has three analytical modules: LinkFinder, LinkInterpreter, and LinkCompare. LinkFinder allows users to search for attributes that are associated with a query attribute, such as mRNA or protein expression signatures of genomic alterations, candidate biomarkers of clinical attributes, and candidate target genes of transcriptional factors, microRNAs, or protein kinases. Scatter plots, box plots, or Kaplan-Meier plots are used to visualize analysis results. Association between the query attribute and individual attributes in the search space can be calculated using an appropriate statistical test depending on the data types of the two attributes. Each query in LinkFinder will return statistical test results for all attributes in the user-defined search space (e.g. all mRNA transcripts or all proteins) and each result can be visualized. To derive biological insights from the association results, the LinkInterpreter module performs enrichment analysis based on Gene Ontology, biological pathways, and network modules, among other functional categories. The LinkCompare module uses visualization functions (interactive Venn diagrams, scatter plots, and sortable heatmaps) and meta-analysis to compare and integrate the association results generated by the LinkFinder module, which supports multi-omics analysis in a single cancer type or pan-cancer analysis.

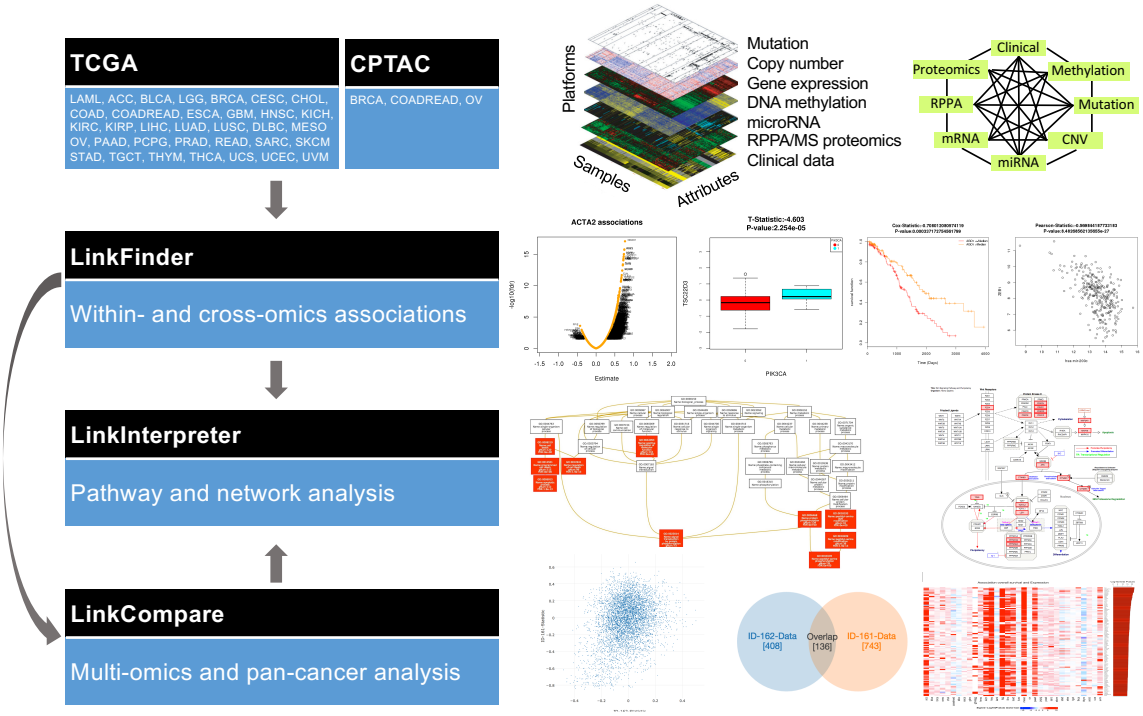


Figure 1. LinkedOmics overview. The data for cancer cohorts are obtained from the TCGA and CPTAC portals. The upper panel shows multi-dimensional omics data and possible pair-wise associations. The user can access the data

and perform directed analysis using three major modules: LinkFinder (performs omics association), LinkInterpreter (pathway enrichment and network analysis), and LinkCompare (multi-omics or pan-cancer analysis). The right panel shows the analysis results obtained from LinkedOmics respective to each module.

1.2 Citation

Suhas Vasaikar, Peter Straub, Jing Wang, Bing Zhang. LinkedOmics: analyzing multi-omics data within and across 32 cancer types (Under Review).

1.3 Get Assistance

For assistance, post your problem in the user forum at https://groups.google.com/forum/#!forum/linkedomics_zhanglab. User can also email admin@Linkedomics at linkedomics.zhanglab@gmail.com.

2. Data Source

2.1 Source

TCGA genomic, epigenomic, and transcriptomic data were downloaded from the TCGA data portal, where the data are available at four different levels. Levels 1 through 4 correspond to raw, processed, segmented/interpreted, and summary data, respectively. To avoid redundant effort, we used pre-processed data from the Firehose Pipeline of the Broad Institute (http://gdac.broadinstitute.org/). Clinical data for the tumors were downloaded from the TCGA data portal (http://cga-data.nci.nih.gov/tcga). The clinical data includes overall survival time, tumor site, age, histological type, lymphatic invasion status, lymph node pathologic status, primary tumor pathologic spread, tumor stage, and radiation therapy status. Molecular subtype, tumor purity, and platinum status data are obtained from the literature.

CPTAC proteomic data include global protein expression data and post-translational modification (PTM) data generated by mass spectrometry (MS)-based shotgun proteomics. These data are available at the raw, mzML, peptide-spectrum match (PSM), and protein levels through the CPTAC data portal (https://cptac-data-portal.georgetown.edu/cptacPublic/). All downloaded datasets were properly normalized and stored in the database.

Table 1. LinkedOmics contains primary tumor data obtained from the TCGA 2016[a] version.

Cancer Cohort	Samples	Unique Samples
Adrenocortical carcinoma (ACC)	92	92
Bladder urothelial carcinoma (BLCA)	412	412
Breast invasive carcinoma (BRCA)	1097	1097
Cervical and endocervical cancers (CESC)	307	307
Cholangiocarcinoma (CHOL)	45	45
Colorectal adenocarcinoma (COADREAD)	629	629
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC)	48	48
Esophageal carcinoma (ESCA)	185	185
Glioblastoma multiforme (GBM)	595	595
Glioma (GBMLGG)	1110	
Head and Neck squamous cell carcinoma (HNSC)	528	528
Kidney Chromophobe (KICH)	113	113
Pan-kidney cohort (KICH+KIRC+KIRP) (KIPAN)	941	
Kidney renal clear cell carcinoma (KIRC)	537	537
Kidney renal papillary cell carcinoma (KIRP)	291	291
Acute Myeloid Leukemia (LAML)	200	200
Brain Lower Grade Glioma (LGG)	515	515
Liver hepatocellular carcinoma (LIHC)	377	377

Lung adenocarcinoma (LUAD)	522	522
Lung squamous cell carcinoma (LUSC)	504	504
Mesothelioma (MESO)	87	87
Ovarian serous cystadenocarcinoma (OV)	591	591
Pancreatic adenocarcinoma (PAAD)	185	185
Pheochromocytoma and Paraganglioma (PCPG)	179	179
Prostate adenocarcinoma (PRAD)	499	499
Sarcoma (SARC)	261	261
Skin Cutaneous Melanoma (SKCM)	470	470
Stomach adenocarcinoma (STAD)	443	443
Stomach and Esophageal carcinoma (STES)	628	
Testicular Germ Cell Tumors (TGCT)	134	134
Thyroid carcinoma (THCA)	503	503
Thymoma (THYM)	124	124
Uterine Corpus Endometrial Carcinoma (UCEC)	548	548
Uterine Carcinosarcoma (UCS)	57	57
Uveal Melanoma (UVM)	80	80
Total 35	13837	11158

[a]The version obtained from firehose updated on 01/28/2016

2.2 "OMICS" Data

- Clinical Data: Includes attributes like age, overall survival, pathological stage (I, II, III, IV), TNM staging, molecular subtype, number of lymph nodes, radiation therapy, tumor purity, and platinum status (only for OV)
- Copy Number (Level: Focal, Gene): Normalized copy number (SNPs) and copy number alterations for aggregated/segmented regions, per sample
- miRNA (Level: Gene): Normalized signals per probe or probe set for each participant's tumor sample
- Mutation (Level: Site, Gene): Mutation calls for each participant
- Methylation (Level: Site, Gene): Calculated beta values mapped to the genome, per sample
- RNAseq (Level: Gene, Isoform): The normalized expression signal of individual genes or isoforms (transcripts), per sample
- Expression Microarray (Level: Gene): Normalized mRNA expression for each gene, per sample
- RPPA (Level: Analyte, Gene): Normalized protein expression for each gene, per sample
- Proteomics (Level: Gene): Average is the log-ratio of sample reporter-ion to a common reference of peptide ions of peptides that match uniquely to the gene. Number of spectra matched to peptides that match uniquely to the gene.
- Phospho-proteomics (Level: Site and Gene): Average log-ratio of sample reporter-ion to a common reference of peptide ions associated with phosphorylated site combinations (CDAP Protein Report). For gene level

data, maximum IQR (inter quartile region) used to select representative site towards gene level.

- Glycoproteomics (Level: Site and Gene): Average log-ratio of sample reporter-ion to a common reference of peptide ions associated with deglycosylated N-glycosylation site combinations (CDAP Protein Report). For gene level data, maximum IQR (inter quartile region) used to select representative site towards gene level.

For more information visit
https://cancergenome.nih.gov/abouttcga/aboutdata/datalevelstypes.

The omics data consist of continuous data (e.g. abundance of miRNA, mRNA, or protein), binary data (e.g. somatic mutation status), categorical data (e.g. tumor histological type), ordered categorical data (e.g. tumor stage), and survival data (e.g. patient disease-free survival time).

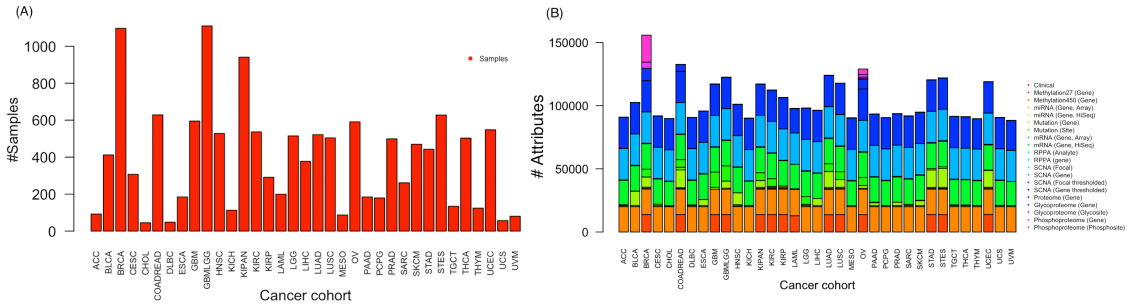


Figure 2. Omics data for 32 cancer cohorts downloaded from TCGA (version 2016_01_28). (A) The bar plot shows the number of cancer samples associated with cancer cohorts. (B) The omics data consist of clinical information, copy number variation, mutation, methylation, miRNA expression, mRNA expression, and protein expression by RPPA. Proteome BRCA from Broad institute, Proteome COADREAD from Vanderbilt University, Proteome OV from JHU-Johns Hopkins University/University of North Carolina, PNNL-Pacific Northwest National Labs. Phosphoproteome BRCA from Broad institute, Phosphoproteome BRCA from PNNL-Pacific Northwest National Labs. Phosphosite BRCA and OV from Broad institute.

2.3 Data Curation and Integration

The downloaded data were curated for NAs (>60%). For gene level data, the minimum number of observations required was 5 or more, while for mutation site level it was 3 or more. The TCGA cancer cohort data were submitted into the server-based “LinkedOmics Database” consisting of data specific tables. For example, breast cancer cohort (BRCA) patient-associated methylation, somatic mutation, mRNA expression, CNV, RPPA, miRNA, and phosphoproteome data are stored in assigned tables. A similar procedure was followed for the other cancer cohorts. Currently, the LinkedOmics database stores and delivers the relevant cancer cohort information as of 2016. The portal will be updated on a regular basis.

3. Quick Start

3.1 Web Interface

The LinkedOmics web interface can be accessed using a guest login or personal login. The personal login saves the queried data.

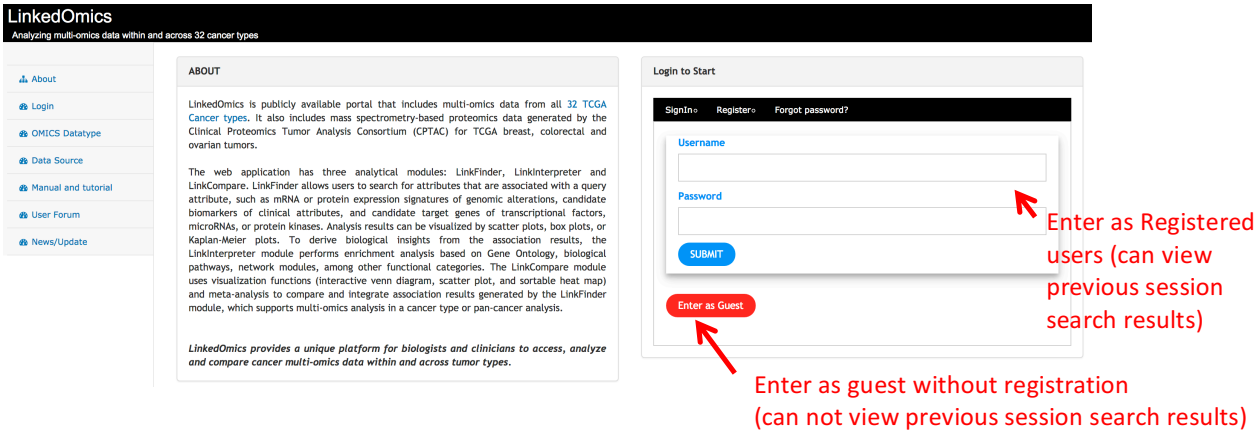


Figure 3. LinkedOmics login page. Users can access the query page through a personal account, which automatically stores the personal queries and results. Users also can login without a personal account but will not be able to view the previous queried results.

Other information can be obtained through the main page.

- About : Linkedomics Overview
- Start : Login or enter as guest to explore
- OMICS Data : Datasets incorporated into portal
- Data Source : Source of the data
- Manual and tutorial : Manual to follow the portal
- User Forum : Submit your queries or comments or discuss
- News/Update : Updates
- Share Your Data : How to share your omics data

3.2 Query Panel

The main page is divided into two panels: navigation and query or output (**Figure 4**). A query (right panel) requires three items: a cancer cohort, search dataset (dataset from which attribute/gene will be selected), an attribute of interest (such as a particular gene or phosphorylation site), and a target dataset (dataset with association will be calculated) for pairwise associations with the attribute of interest.

Follow the steps,

1. First select the cancer cohort (say TCGA Breast Cancer, BRCA)
2. Select search dataset containing the attribute of interest (say the gene level mutation dataset).

3. Optional (Select population with particular characteristics. For example, select only “ER positive” patients for further analysis. Users can add more attributes with “OR” logic criteria).
4. Select attribute/gene of interest (say the TP53 gene).
5. Select target dataset with which possible pair-wise association analyses will be calculated (like RNAseq expression).
6. Select statistical method among LinkedOmics suggested options (For example, the relation between *TP53* gene mutations and mRNA expression of BRCA cohort patients can be studied by Student t-test or Wilcoxon test).

After selecting the appropriate statistical test, LinkFinder performs the computation on-the-fly on the server side and outputs the result in table format (**Figure 5**). [Note: The omics dataset panel for each cancer cohort was named based on features or annotations. The specific features and user submission requirements are given in **Section 5: Data Sharing and Annotation**].

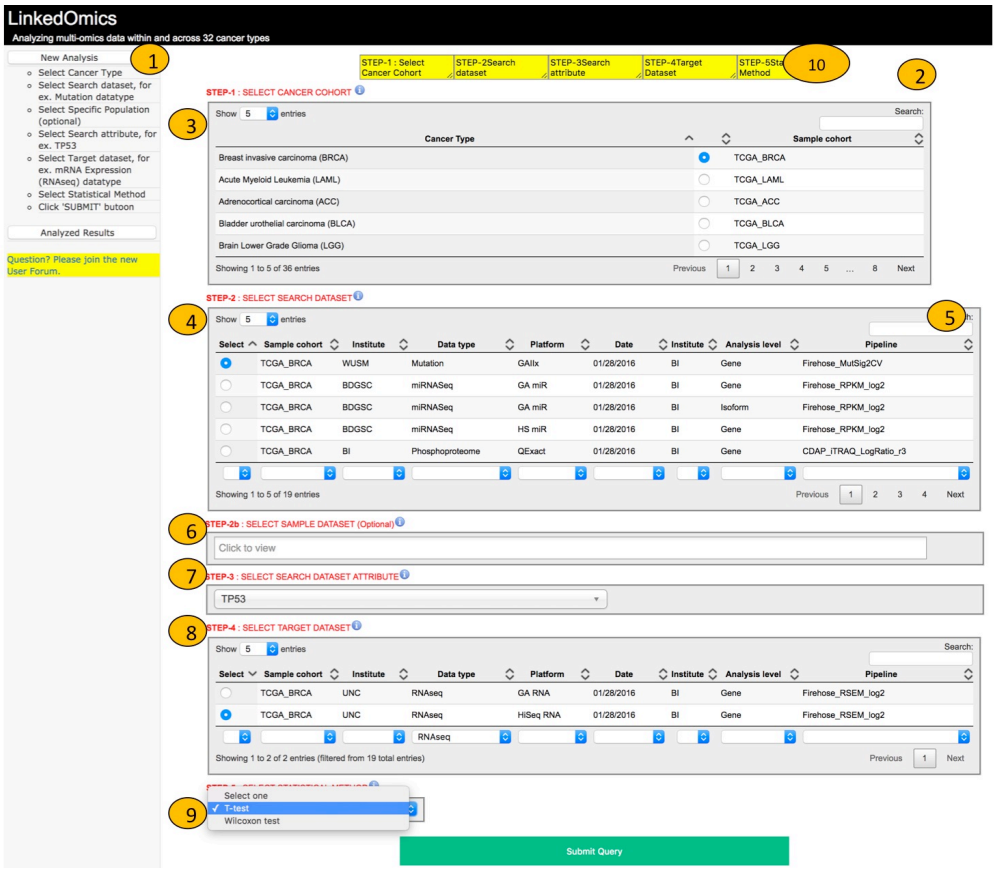


Figure 4. The query panel. The query page is divided into two panels: navigation (1) and query or output (2). From the right panel (query panel), the user can select the cancer cohort from 32 cancer types (35 cancer cohorts) (3). After selecting the cancer cohort, a search dataset panel appears (4) showing omics data available for the given cancer cohort. The user can search for a

specific dataset using the search bar (5). The user can also restrict the search to a sample population with particular characteristics (**optional**) (6). A search attribute must be selected after choosing the specific search dataset of interest (7). For pair-wise associations, the target dataset is selected from the target dataset panel (8). Based on the data types of the target dataset and the search attribute from the search dataset, an appropriate statistical method is selected (9). The user can visualize the stepwise selection progress at the top of the page (10).

We built a "data cart" system that allows users to store multiple sets of association results during analysis. Selecting the particular query from the data cart (or query summary options) will show the result.

3.2.1 LinkFinder Module

The LinkFinder module allows users to search for statistically significant attributes that are associated with a query attribute. The search can be limited to a subspace (e.g. transcript abundance data, protein abundance data, or protein phosphorylation data). Association between the query attribute and individual attributes in the search space is calculated using an appropriate statistical test depending on the data types of the two attributes (see section **3.3 Statistical analysis**). Association results are returned to users in a column-sortable table (**Figure 5**). All of the analysis is performed on-the-fly.

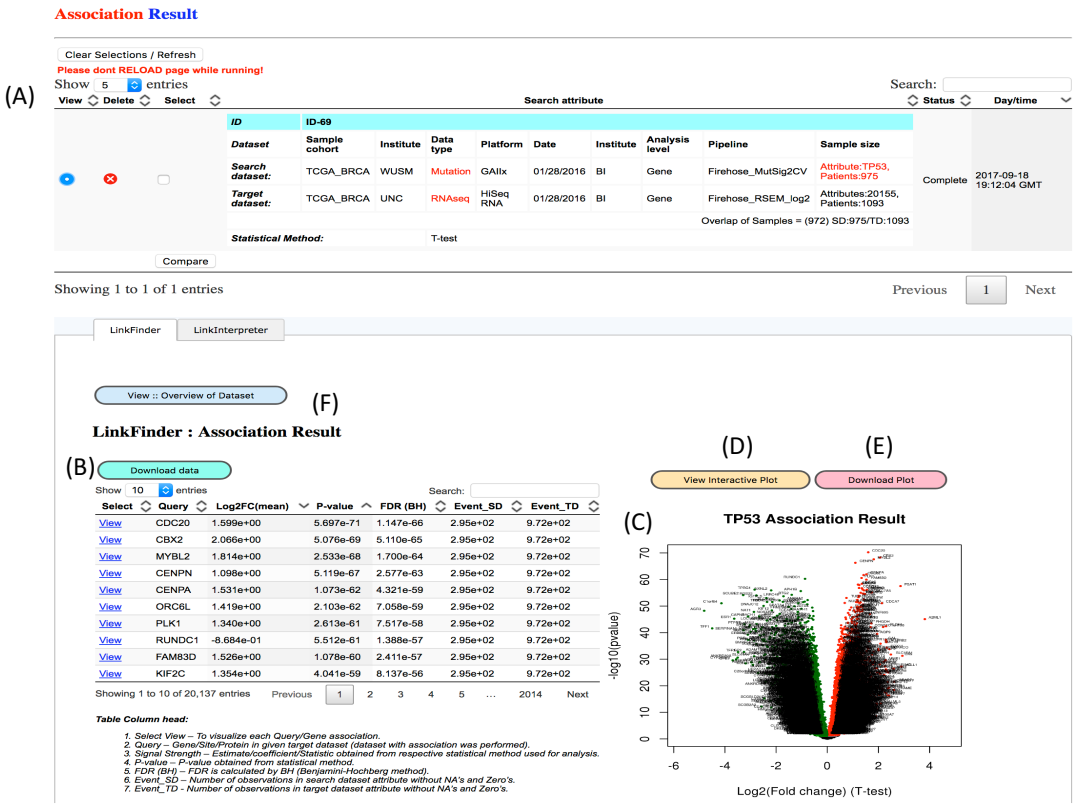


Figure 5. Example output from the LinkFinder module. (A) Tabular view showing the association between *TP53* mutation and mRNA expression in breast cancer. (B) Statistically significant correlation results using a T-test is shown in the lower panel. (C) The volcano plot showing the Log2(fold change) vs. –log10(p-value) obtained from analysis. The plot can be visualized interactively (D) or downloaded (E). The overall results summary is given in "Overview of datasets" (F).

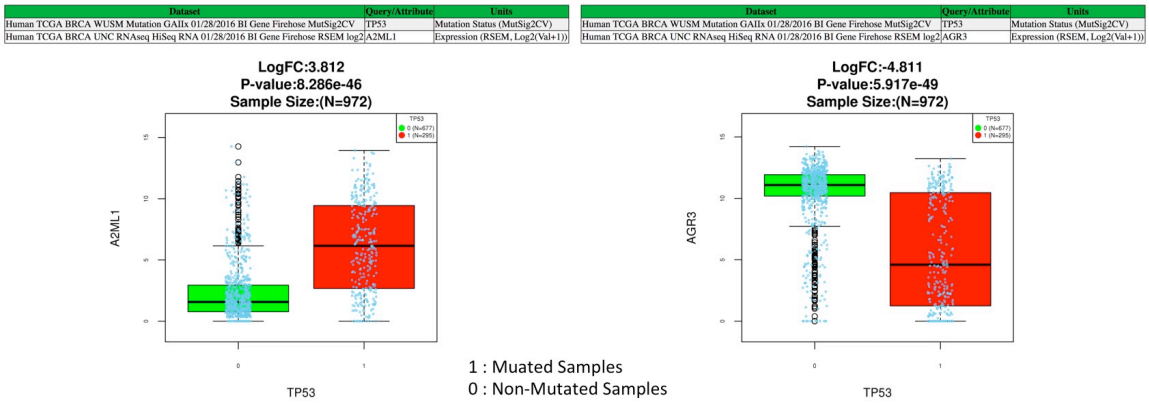


Figure 6. *TP53* mutation in breast cancer. *A2ML1* and *AGR3* gene expression in either direction is found to be significant in terms of log2(Fold change) and P-value. Positive correlation/up-regulation of *A2ML1* gene expression (A) and negative correlation/down-regulation of *AGR3* gene expression in *TP53*-mutant samples are shown (B).

The visualization of individual associations may differ depending on the data types and selected statistical analyses of the two associated attributes (**Figure 7A-C**).

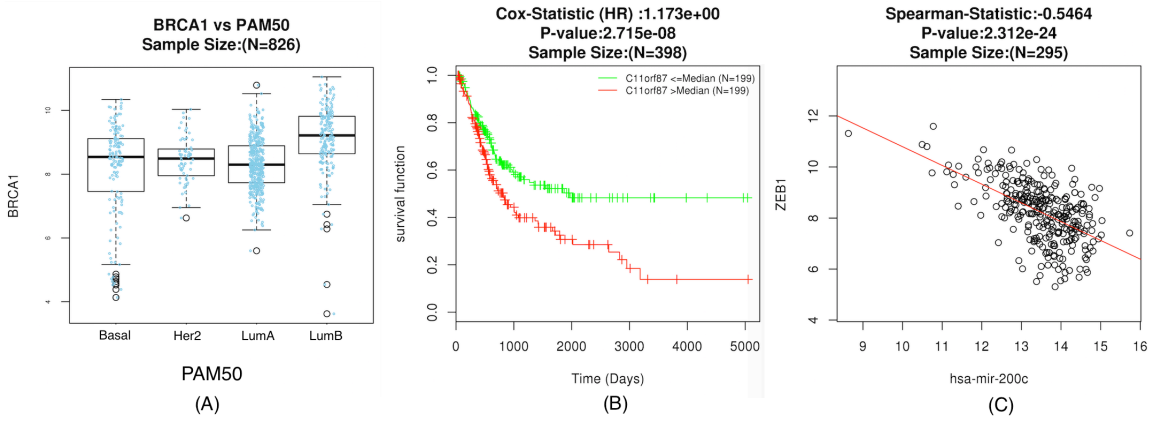


Figure 7. LinkFinder association results. (A) Box plot showing the relationship between *BRCA1* expression and PAM50 subtypes in breast cancer. (B) Kaplan–Meier plot comparing patient survival time between samples with high- or low-level of *C11orf87* expression in Bladder cancer (BLCA). (C) Scatter plot showing

the relation between miRNA has-mir-200c and *ZEB1* gene expression in COADREAD.

3.2.2 LinkInterpreter Module

LinkInterpreter performs gene enrichment analysis on genes of interest selected from the LinkFinder result. Clicking on the "LinkInterpreter" tab, which is next to the tab "LinkFinder", can access the enrichment panel. The panel shows a dropdown menu to select enrichment methods: ORA (over-representation analysis) or GSEA (gene set enrichment analysis) (**Figure 8A-B**). Selecting either method directs the user to the respective enrichment analysis parameter selection panel. **Figure 8C** shows the parameter selection panel for the ORA method of enrichment. The functional database is selected from the dropdown menu. Choices are GO Analysis (Biological process), GO Analysis (Cellular Component), GO Analysis (Molecular Function), KEGG pathway, Reactome pathway, Wiki Pathway, Kinase target enrichment, miRNA target enrichment, transcription target enrichment, and PPI network enrichment (**Figure 8F**). Significance criteria can be selected as FDR (False discovery rate [BH]), p-value, or Top (**Figure 8D**). Further, the user can select positively correlated genes or negatively correlated genes with respect to the question under study (menu Select sign (or direction)). The direction is obtained from the test statistic. The user can also select the "Both" option. Significance for FDR or P-value can be chosen as per the user-set threshold (default 0.05). Top genes also can be selected with the user-defined number (default 100). The GSEA-based analysis can be performed by selecting the GSEA method selection (**Figure 8E**). The functional databases are shown in **Figure 8F**.

The submit button directs the user to a new page. The new page has 2 tabs: "View Filtered Data" and "View Enrichment Results" (**Figure 9**). The Enrichment results tab displays the results obtained using the chosen functional database (here *Gene ontology for biological process*). The summary of results can be accessed from the tab "*Summary*" (**Figure 9A**). Detailed results for gene ID mapping (**Figure 9B**), GOSlim (**Figure 9C**), and enrichment analysis (**Figure 9D**) can be found using the other three tabs. Finally, the user can download the results from "Result Download" link on the summary page (**Figure 9E**). Users should go through the summary page for more information.

The "View Filtered Data" tab displays the genes or attributes selected based on the user-defined criteria (**Figure 10A**). A table is shown with attributes and relevant metric (FDR or P-value and Statistic). Further, the user can select the button "IdeogramViewer" for the chromosomal view (**Figure 10B**). In the chromosomal view, genes are highlighted with red bars. Users can click on each red bar to view the gene of interest.

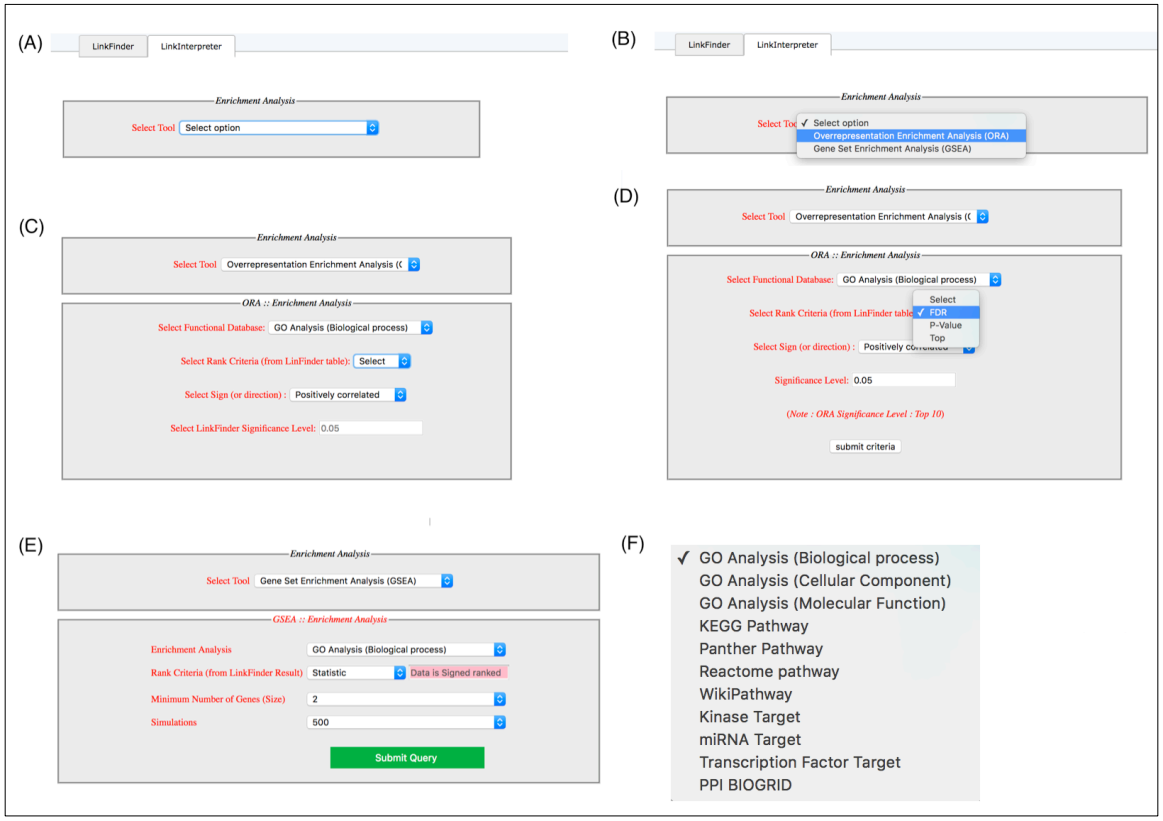


Figure 8. LinkInterpreter overview. LinkInterpreter facilitates enrichment-based analysis using ORA (over-representation analysis) or GSEA (gene set enrichment analysis) (A, B). In the ORA panel, the user can select a functional database (C) and a significant gene list based on FDR, p-value, or top number of significant genes (D). In the GSEA panel, the user can select rank criteria, minimum number of genes in the list, and number of simulations (E). The functional databases are shown in (F).

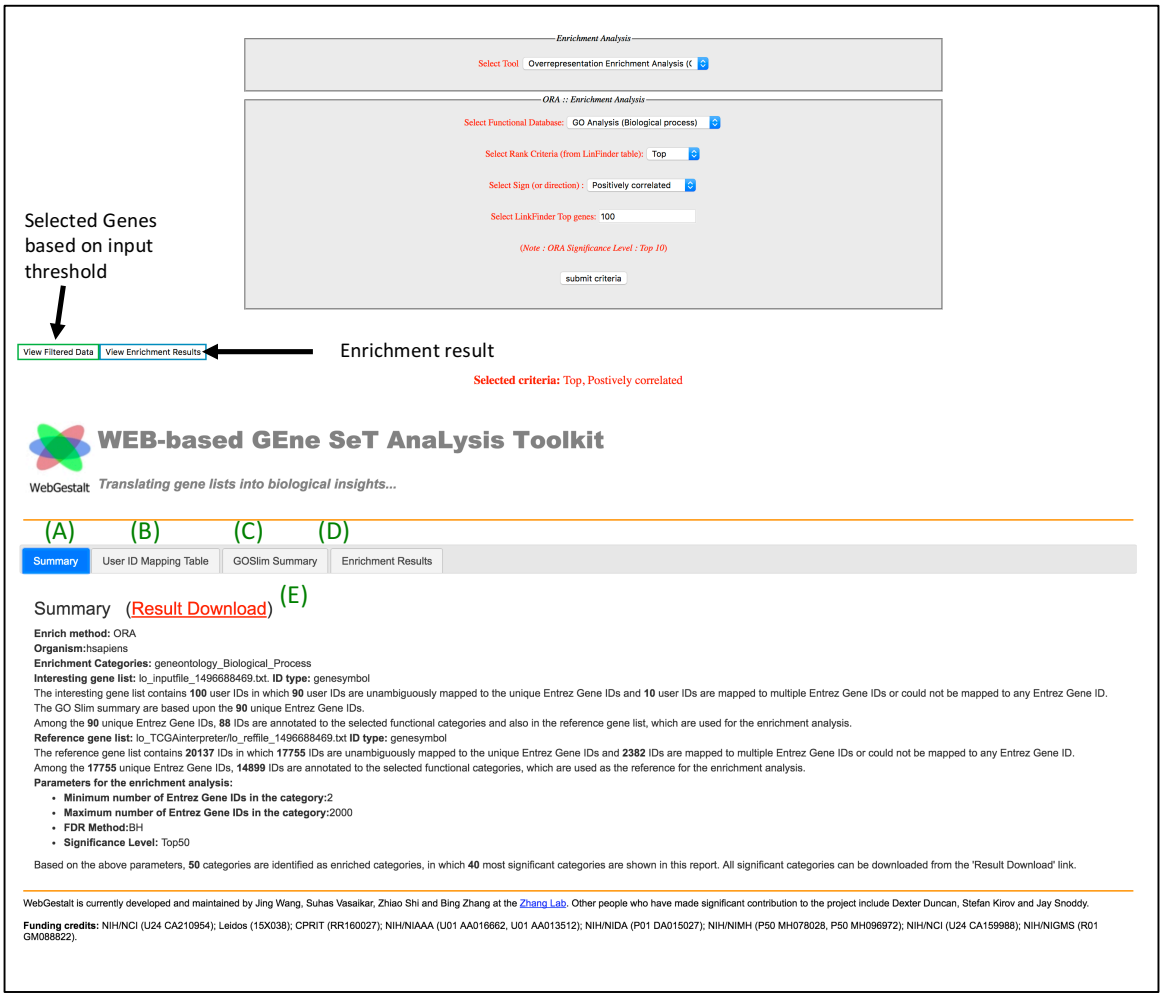


Figure 9. Selection of the ORA method of enrichment using the functional database "GO analysis (Biological Process)". The top 100 genes are selected for genes positively correlated with the *TP53* mutation in the breast cancer cohort.

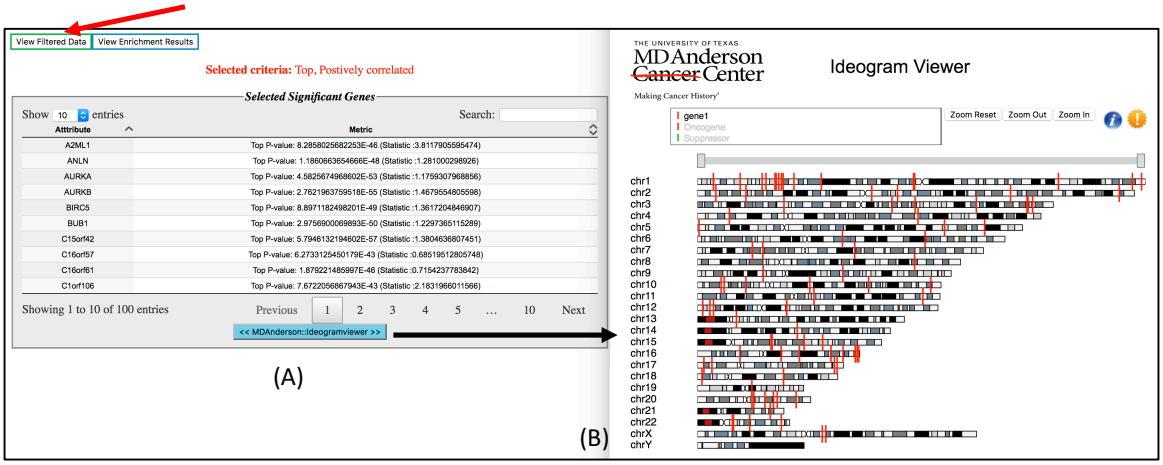


Figure 10. LinkInterpreter results. (A) The "View Filtered Data" tab shows the filtered gene list from the LinkFinder result. (B) The selected genes and their chromosomal locations are shown using "IdeogramViewer" (http://bioinformatics.mdanderson.org/main/IdeogramViewer:Overview).

The GOSlim result (biological process, cellular component, and molecular function categories) is shown (**Figure 11A**). The result is obtained for the *TP53* mutation-associated significant gene enrichment analysis (FDR < 0.05). The visualization of the Gene Ontology (GO) for biological process is shown using a parent-child relationship (**Figure 11B**). The FDR and gene size obtained for each enriched process is depicted in the graph. On the right side, the genes associated with each enriched process is shown in tabular format. Users can search for genes or ontology in the "search panel". *TP53* mutation correlated with increased cell cycle activity.

"GSEA" method of enrichment selection displays the parameter selection panel. Users can select the functional database as described in **Figure 8**. The test-statistic or signed p-value is used to perform the analysis. Users can select a minimum number of genes and number of simulations for their analysis (**Figure 12A**). The enrichment results table is shown in the right panel. The Gene Ontology database-based enrichment finds significant biological processes, which can be visualized in the ontology viewer (**Figure 12B**). The enriched biological processes corresponding to positively correlated genes are shown in red blocks, while negatively correlated genes are shown in blue blocks. Users can explore the data dynamically. The biological processes affected by *TP53* mutation are shown here.

The enrichment analysis is carried out by our popular tool WebGestalt (http://www.webgestalt.org). The WebGestalt API provides a gene level analytical tools platform for the interpretation (including pathway and ontology visualization) of LinkFinder results.

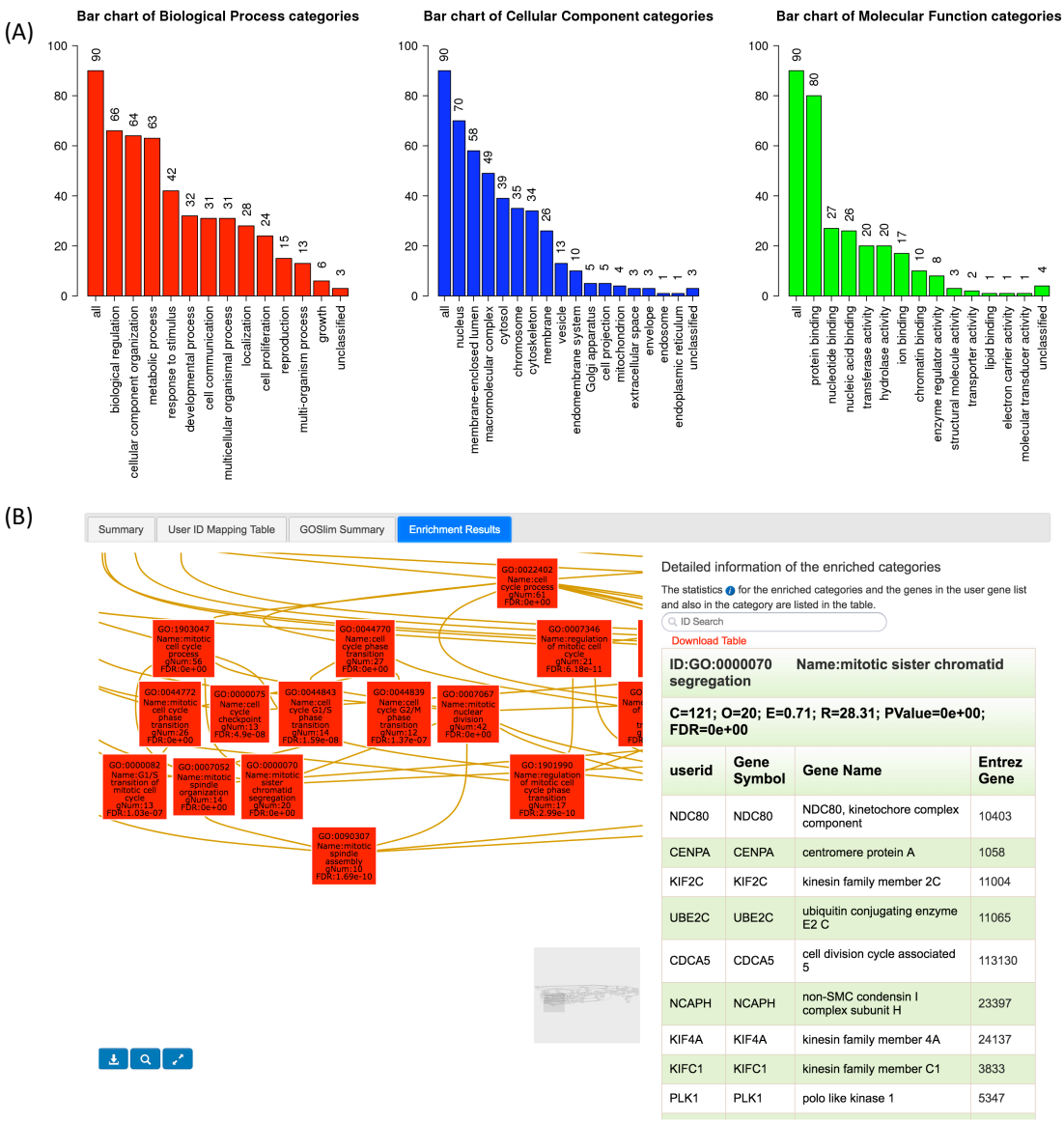


Figure 11. ORA GO results for the *TP53* mutation associated with gene expression. (A) The GOSlim result (biological process, cellular component, and molecular function categories) obtained using the ORA method is shown. (B) Gene ontology-based enriched biological processes are shown using parent-child relationships. The red blocks show significant results obtained (FDR and gene number is given). On the right side, the genes associated with each enriched process is shown. The user can download the image or data by clicking on the download button.

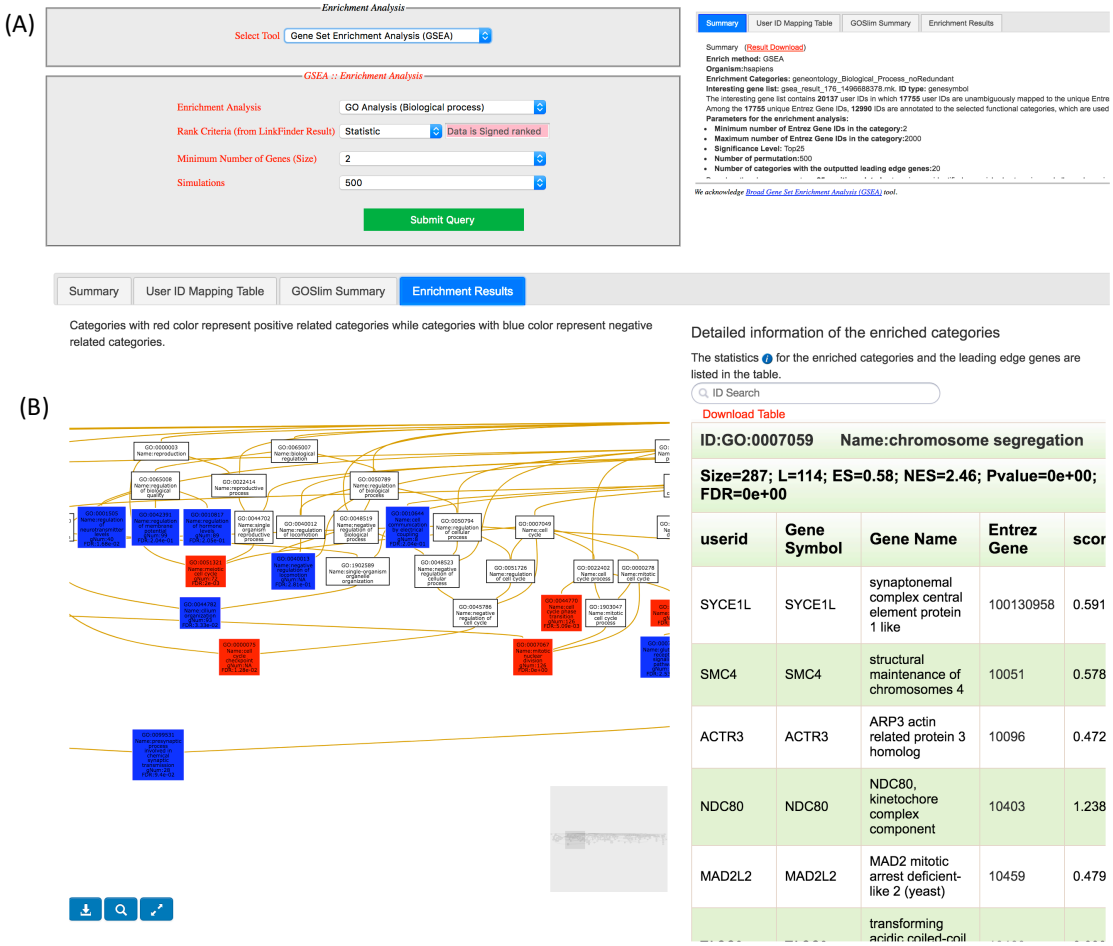


Figure 12. Example of GSEA enrichment results. (A) Selection of the "GSEA" method of enrichment using the functional database "GO analysis (Biological Process)". (B) Biological processes affected by *TP53* mutation in the breast cancer cohort are shown.

3.2.3 LinkCompare Module

LinkCompare performs comparisons between multiple association results generated from LinkFinder. Possible comparisons include those on the same dataset (e.g. proteins associated with *KRAS* mutation vs *BRAF* mutation), or with the same query attribute on datasets from different omics platforms (e.g. genes associated with miR21 in the colorectal cancer RNA-Seq vs shotgun proteomics datasets), tumor types (e.g. genes associated with disease-free survival in colorectal, breast, or ovarian cancer), or tumor subtypes (e.g. proteins associated with AKT phosphorylation in colorectal tumors with or without $KRAS^{G12D}$ mutation). To easily interpret the results, users can select visualization tools to facilitate the comparison.

The LinkCompare module combines the confidence obtained from each omics-associated statistical analysis and performs meta-analysis. Users should refer to our Manuscript for more information. The meta-analysis is performed using the

Stouffer method (Stouffer, 1949). The metaP package is used to obtain the result (Dewey, 2017). The result obtained using the Stouffer method is displayed on the web portal in the tabular format. Here we have shown the comparison between *RB1* mutation on RNA expression in the breast cancer cohort (BRCA) and bladder cancer cohort (BLCA). The Linkfinder-based *RB1* mutation association with RNA expression is shown in **Figure 13A**. To select the comparison, users should click on the checkbox in the Select panel on each respective query result to compare (**Figure 13B**) and then click the "Compare" button on the bottom of the table. After clicking the button, a new panel appears at the bottom, which performs the comparison on-the-fly. It usually takes ~40-60s for two query datasets, and ~60-90s for three datasets. The compare module performs meta-analysis and returns a table and figures (**Figure 13C**). The result output consists of

- Table (**Figure 13**, which shows the LinkFinder result for each selected query type)
    - **Figure 14E**: The attributes such as genes observed in each LinkFinder results are shown.
    - **Figure 14F**: We displayed the query results with specific ID-types. Users should refer to each ID-type given in the upper selection panel. The statistic, P-value, and FDR by BH for each query result (or ID-type) are shown in the table.
    - **Figure 14G**: The meta-analysis result obtained using the Stouffer method such as meta statistic (sumz_stat), meta P-value (sumz_P), and meta FDR (sumz_FDR) are displayed in the table.
- Download Table (**Figure 14H**, the user can download the result in text format).
- Scatter Plot (**Figure 14I**, the user can visualize the comparison between two omics query comparisons using a scatter plot).
- Venn Plot (**Figure 14J**, the user can visualize the comparison between two omics query comparisons using a Venn diagram).

Further, the user can perform an enrichment analysis (pathway, transcription factor, or kinase targets) on meta-analysis obtained results (**Figure 14K**).

Please Note:

- Only the same level data type results are comparable. For example, users can compare "gene level" data with "gene level" but not with other data type levels such as "site level", "analyte level", "focal level", "miRNA level," and "phosphosite level".
- When a user selects a query result for comparison, the portal automatically disables other data type level options.

Query (Gene/Protein/site)	ID-2-SignalStrength	ID-2-Pvalue	ID-2-FDR	ID-1-SignalStrength	ID-1-Pvalue	ID-1-FDR	Sumz_stat	Sumz_P	Sumz_FDR
CDKN2A	9.687e-01	4.559e-02	3.501e-01	4.438e+00	2.517e-49	5.046e-45	1.185e+01	2.075e-32	4.151e-28
MCM6	8.425e-01	1.438e-05	1.187e-03	9.707e-01	8.714e-26	4.622e-22	1.049e+01	9.454e-26	9.456e-22
WDR76	8.144e-01	2.659e-05	1.983e-03	1.249e+00	9.223e-26	4.622e-22	1.039e+01	2.732e-25	1.822e-21
DONSON	1.084e+00	2.449e-06	2.771e-04	1.072e+00	7.544e-23	1.890e-19	1.029e+01	7.794e-25	3.898e-21
MELK	1.226e+00	1.781e-05	1.441e-03	1.187e+00	5.171e-24	1.481e-20	1.018e+01	2.428e-24	9.713e-21
CHAF1B	8.333e-01	3.538e-06	3.711e-04	1.034e+00	7.174e-22	1.438e-18	1.007e+01	7.137e-24	2.379e-20
MCM2	9.677e-01	1.166e-06	1.535e-04	9.832e-01	5.717e-20	7.969e-17	9.907e+00	3.875e-23	1.107e-19
CDKN2BAS	1.021e+00	1.520e-02	1.969e-01	2.606e+00	2.775e-30	2.782e-26	9.803e+00	1.095e-22	2.433e-19
TYMS	1.033e+00	8.570e-06	7.739e-04	1.240e+00	4.139e-21	7.544e-18	9.814e+00	9.776e-23	2.433e-19
E2F1	1.066e+00	1.027e-06	1.389e-04	1.160e+00	1.686e-18	1.536e-15	9.661e+00	4.411e-22	8.823e-19

Figure 13. LinkCompare example. (A) Linkfinder-based *RB1* mutation correlation on mRNA expression in two different cancer cohorts. (B) The comparison between the omics-query dataset performed by selecting the checkbox in the select panel in each LinkFinder query result. Click on the "Compare" button to perform the comparison between the selected query datasets. (C) The meta-analysis is performed and output shown in tabular format. The result can be visualized using a scatter plot or Venn diagram to compare significant attributes across the platforms. Heatmaps are used to compare more than two sets of association results (see **Figure 15**).

Comparison of two OMICS data analysis queries can be visualized as a scatter plot or a Venn diagram (**Figure 14**). A scatter plot allows a user to view the query-based statistic, p-value, or FDR for genes in the corresponding LinkFinder-based result datasets (**Figure 14A**). Whereas, a Venn diagram allows a user to select genes with specific threshold criteria, such as p-value or FDR <0.05 (**Figure 14B**). The input panel accepts the threshold significance, and p-value or FDR is selected using the corresponding button. The lower panel shows the Venn diagram with an overlapping region. A user can access genes corresponding to each region by clicking on the lower box. The ID represents the LinkFinder-based query result. Users should look at the ID before making any interpretation. The result is shown in tabular format in the lower panel.

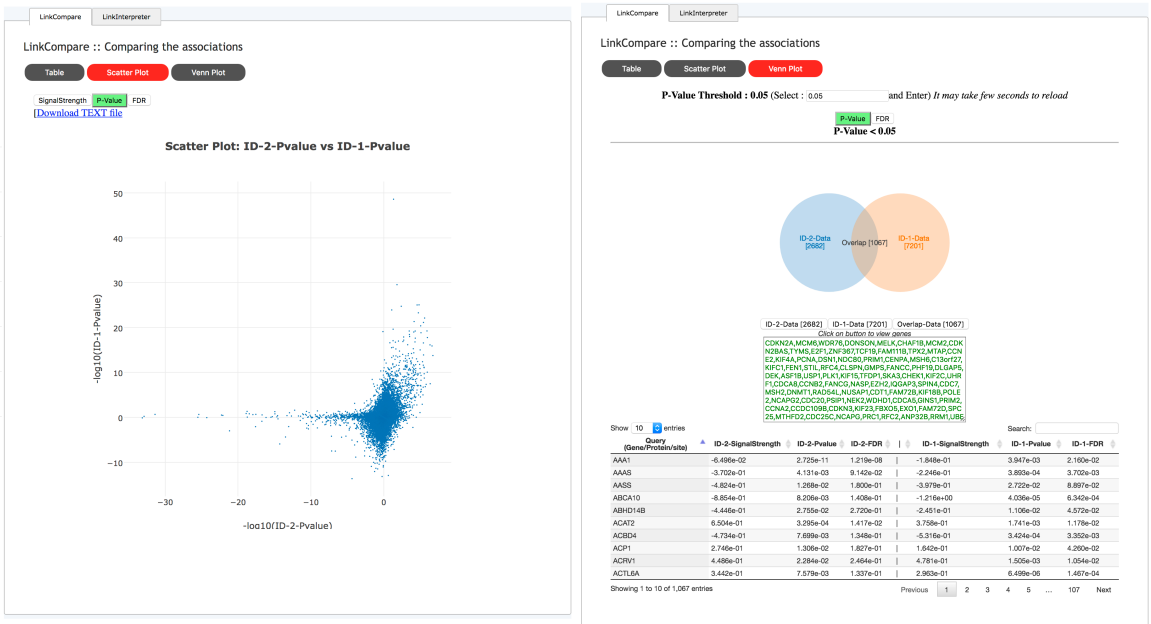
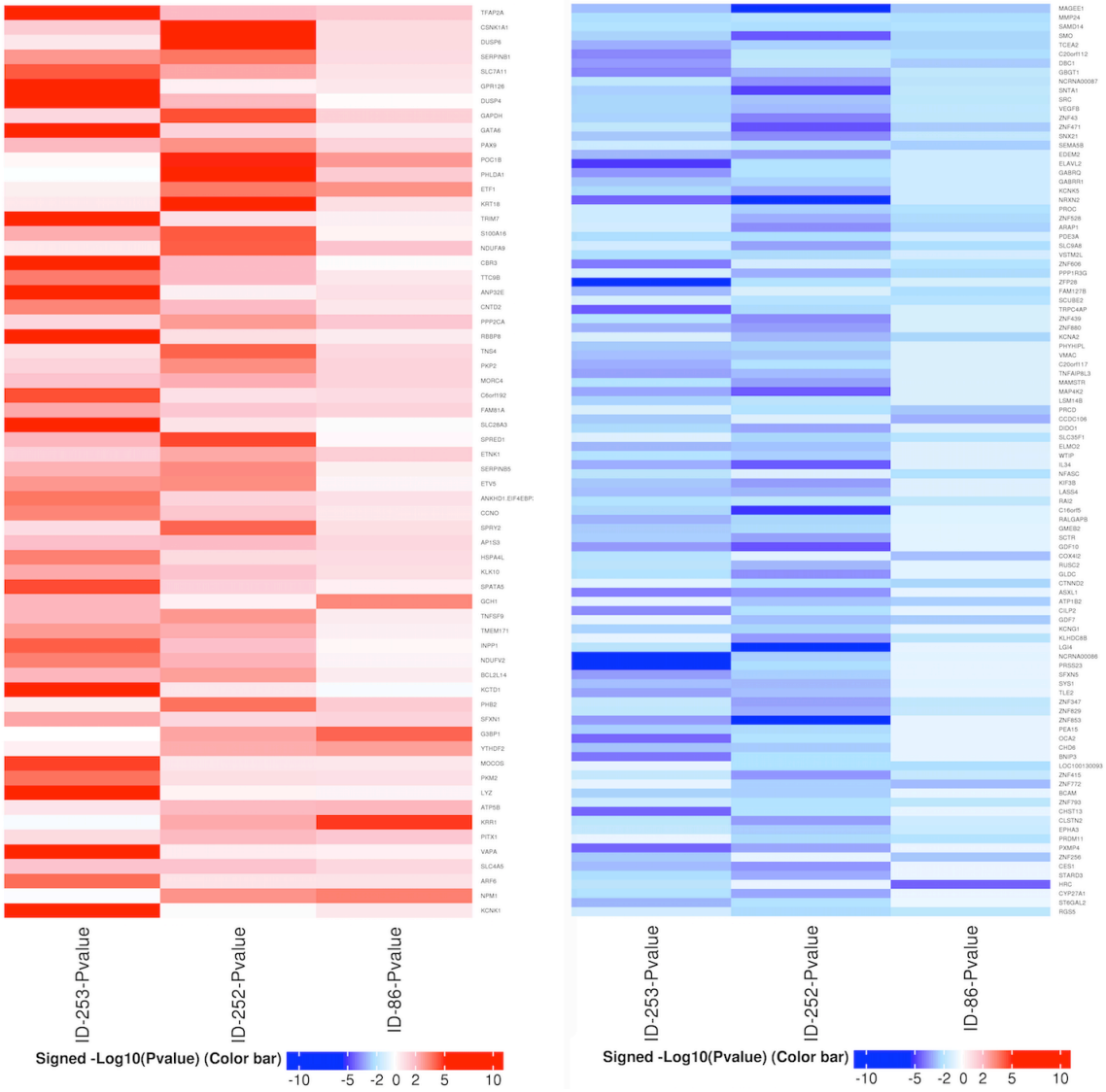


Figure 14. Examples of LinkCompare plots. (A) LinkCompare-based result can be visualized as a scatter plot or (B) Venn diagram. Users can interactively view the attributes and corresponding statistical significance in the graph.

For comparison of more than three OMICS data analyses, the result can be visualized as a heatmap (**Figure 15**). The positively correlated genes are shown in the red spectrum color bar (heatmap in **Figure 15A** and bar plot in **Figure 16A**), while negatively correlated genes are shown in the blue spectrum color bar (heatmap in **Figure 15B** and bar plot in **Figure 16B**).



(A) (B)

Figure 15. (A) Top 100 positively and (B) negatively correlated genes among three omics query results. The scale represents the meta-log10(p value) and the sign is obtained from the meta-statistic.

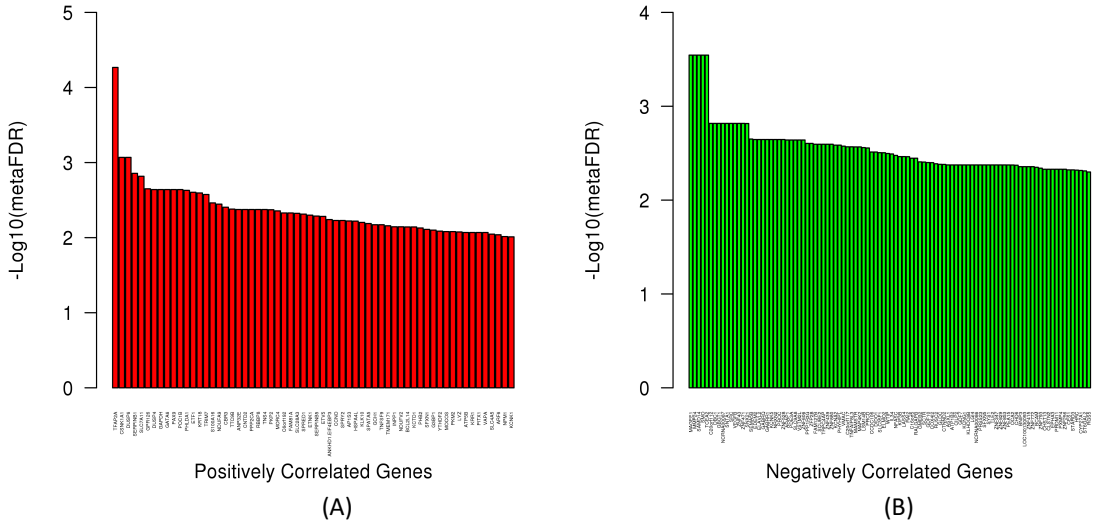


Figure 16. (A) Top 100 positively and (B) negatively correlated genes among three omics query results. The scale represents the meta-log10(p value) and the sign is obtained from the meta-statistic.

3.3 Statistical Analysis

To allow statistical evaluation of the association between different types of data, we have implemented a comprehensive collection of statistical tests such as Pearson's correlation coefficient, Spearman's rank correlation, Students T-test, Wilcoxon test, Analysis of Variance (ANOVA), Kruskal-Wallis analysis, Chi-square test, Fisher's exact test, and Cox's regression analysis. All of these tests are performed using the open source R statistical computing environment. Multiple P-value correction is performed using the Benjamini and Hochberg method (Benjamini and Hochberg 1995).

3.4 Meta-analysis

Meta-analysis is performed using the metaP R package to combine multiple p-values (Dewey 2017). Bootstrapping is used to compare the obtained meta-analysis result with a random meta-analysis result. FDR (false discovery rate) is calculated by combining the real and random meta-analysis results. For more information, users should refer to our manuscript.

3.5 Visualization Methods

We have developed intuitive visualization tools for effectively communicating complex results to a broad audience, such as interactive scatter plots, Venn diagrams, Kaplan-Meier plots, box plots, and heatmaps. The plots can be easily downloaded for further analysis and publication purposes.

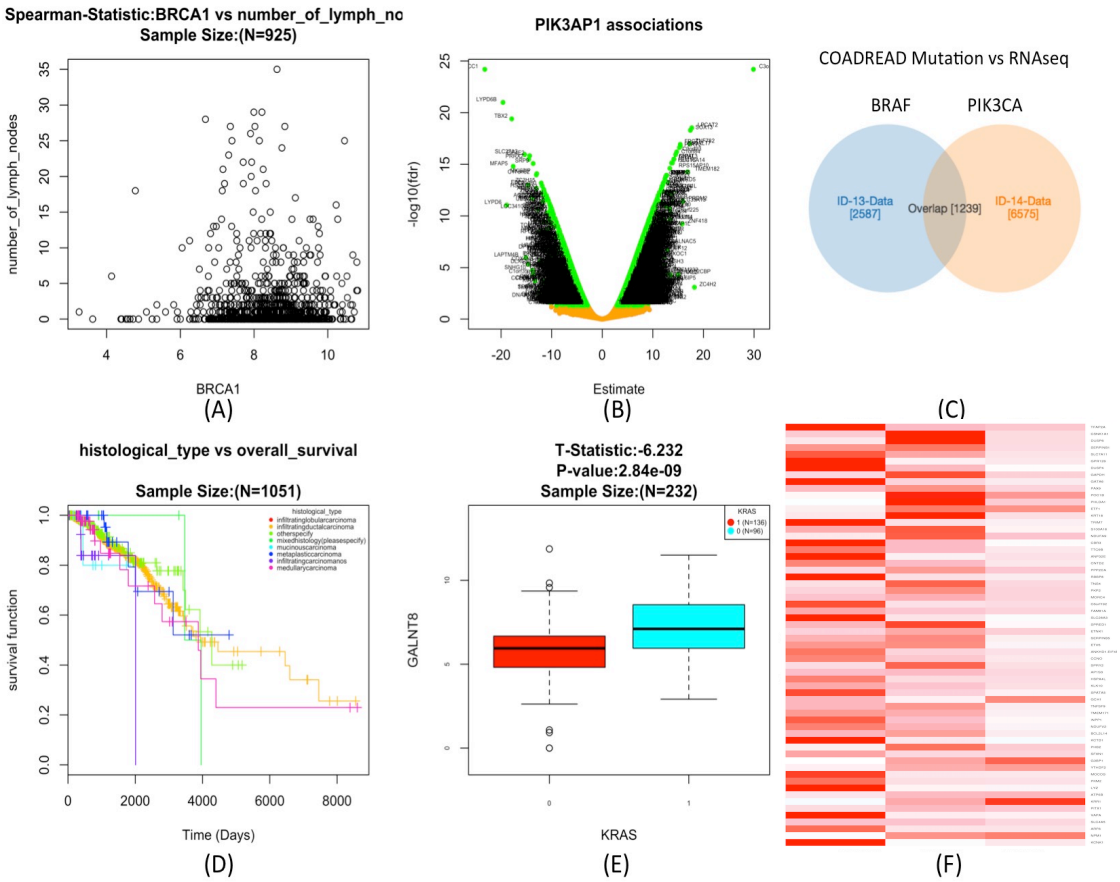


Figure 17. Visualization options. Users can download the result from the output as a (A) scatter plot, (B) volcano plot, (C) Venn diagram, (D) Kaplan-Meier plot, (E) box plot or (F) heatmap.

4. Case Study

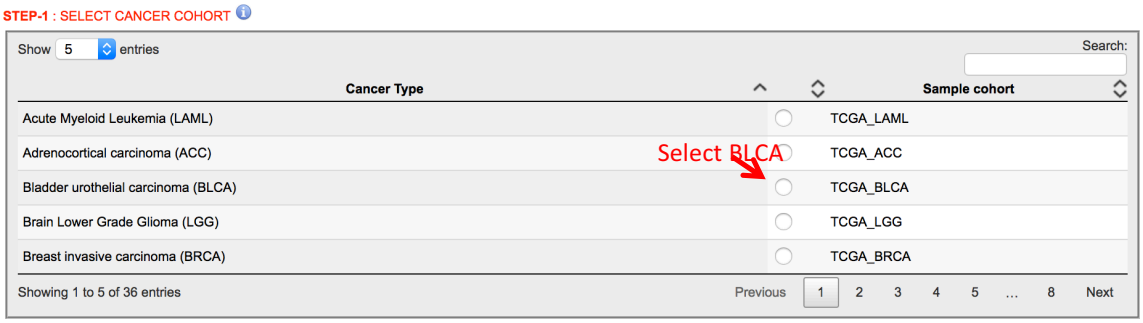
4.1 *RB1* mutation in Bladder urothelial carcinoma (BLCA) and its impact on gene expression

RB1 plays an important role in proliferation, differentiation, and senescence. The protein encoded by *RB1* is a tumor suppressor that acts as a negative regulator of the cell cycle (RefSeq, Jul 2008). The genetic mutation in *RB1* is known to cause tumors in the urinary bladder. Loss of function due to mutation in *RB1* causes the loss of inhibition of the E2F family of transcription factors. This leads to uncontrolled proliferation of epithelial cells. With reference to known literature, we performed a similar analysis of the TCGA Bladder urothelial carcinoma; BLCA cohort.

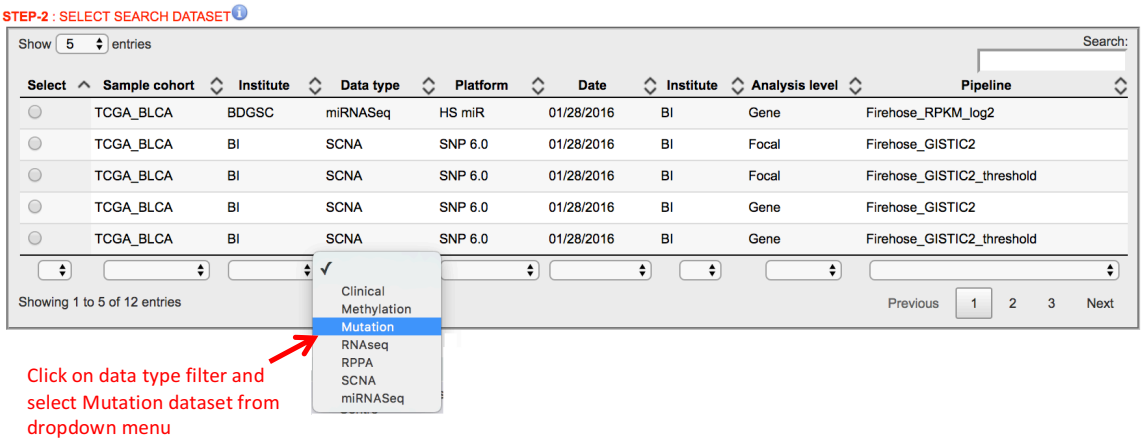
Q. Which genes (RNA expression) are significantly associated with mutation in the *RB1* gene in the Bladder urothelial carcinoma (BLCA) cohort?

Follow the steps:

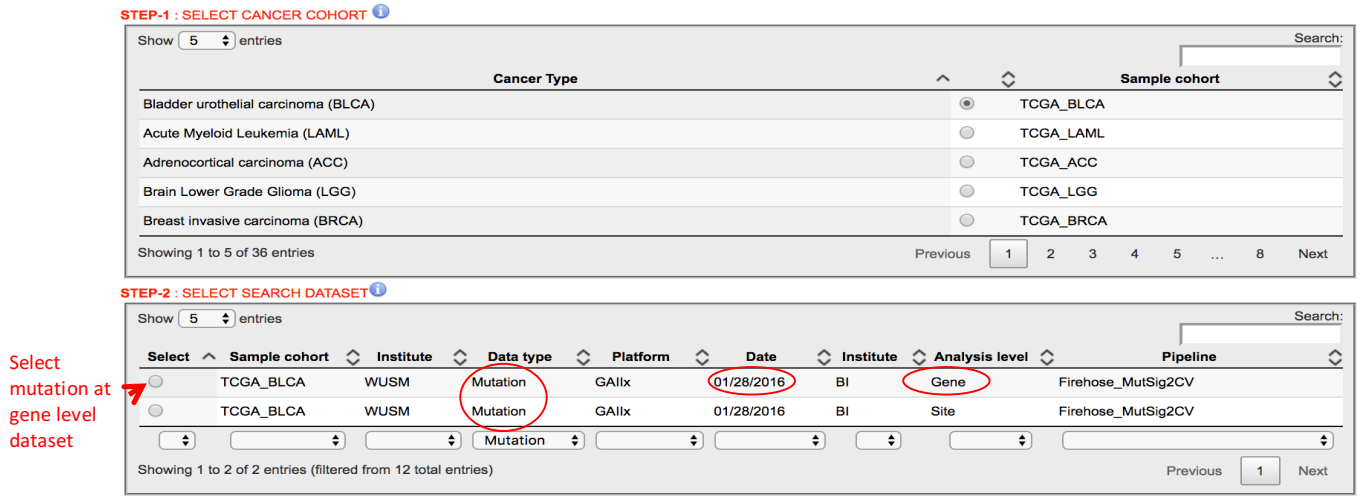
1. Select the cancer cohort “Bladder urothelial carcinoma; BLCA”.



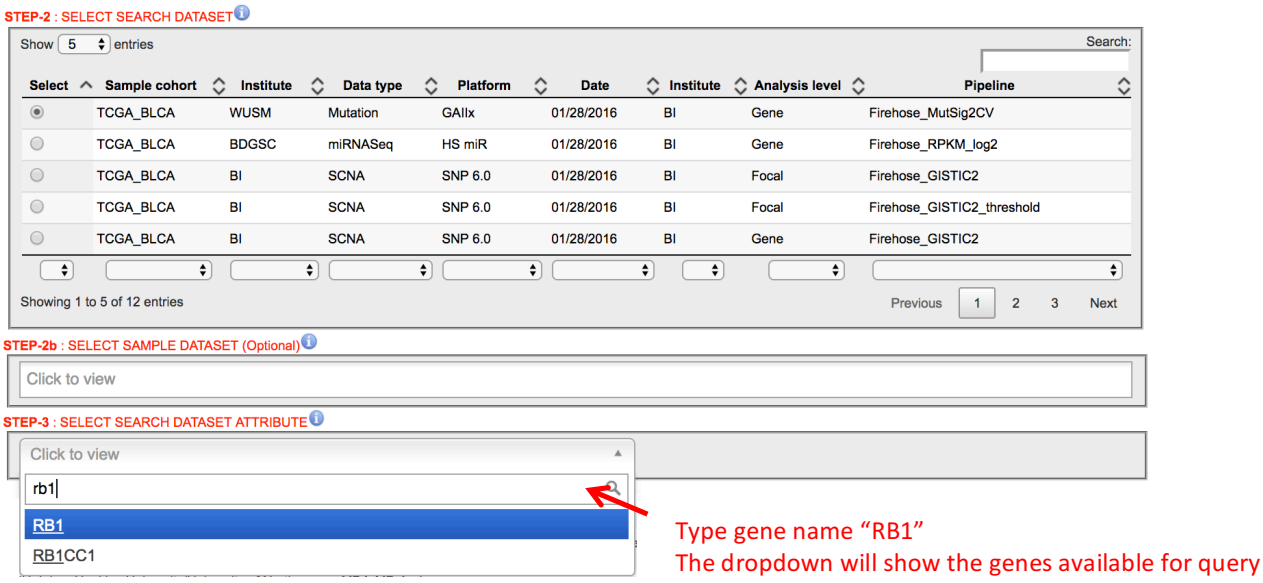
2. Selecting the cancer cohort displays the “Search dataset” panel. Available data types are shown in the “Data type” column. Click on the filter under the “Data type” column and select the “Mutation” data type from the drop-down menu.



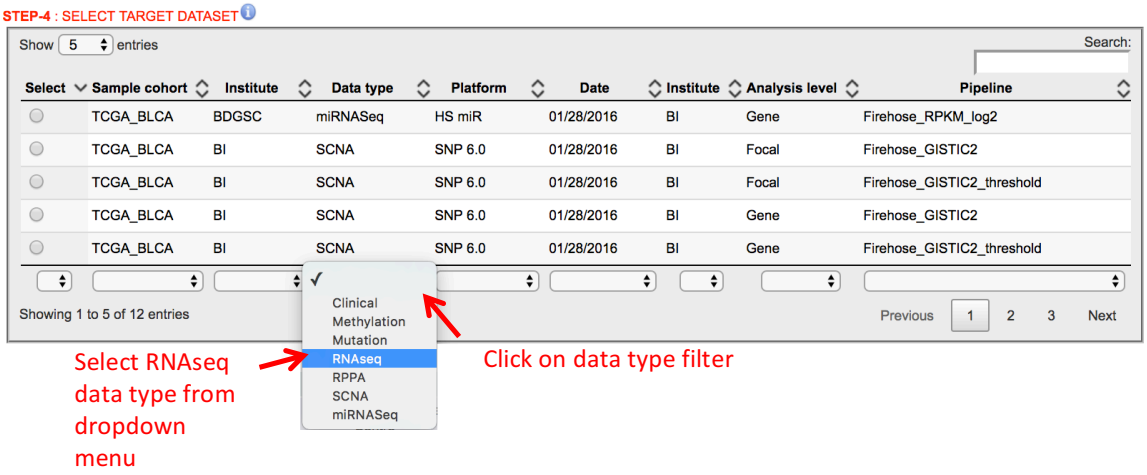
3. Select the gene level “Mutation” dataset, which was obtained from the Broad Firehose with release date “01/28/2016”.



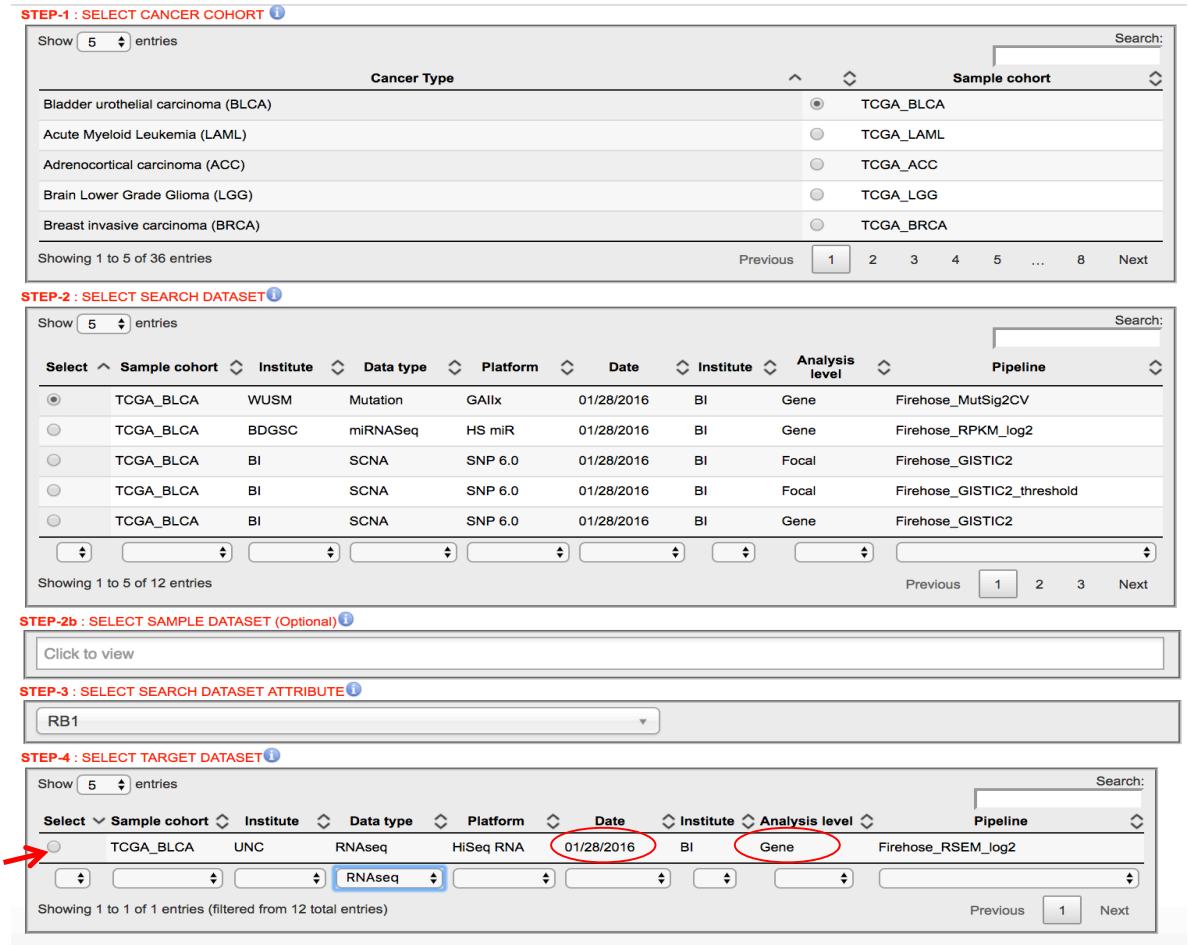
4. After selecting the “Mutation” data type, the “Select Search Dataset Attribute” panel will appear. Type the gene name “RB1” and the dropdown menu will show available attributes for the selected “Mutation” data type. Click on “RB1” in the dropdown menu.



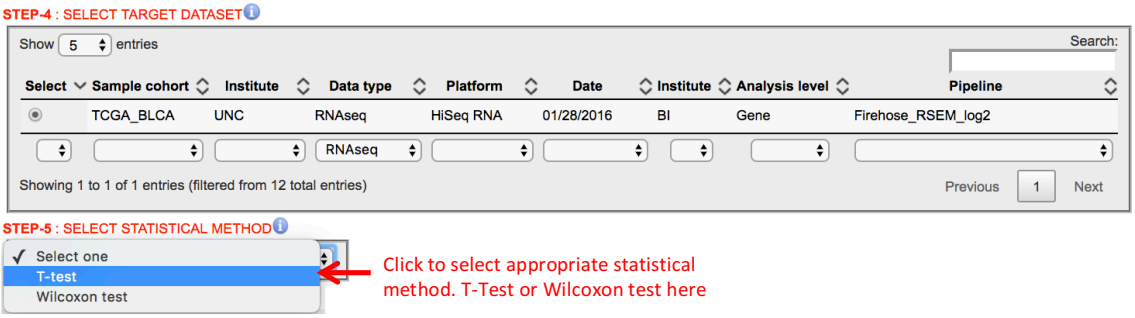
5. Click on Data type filter in the “Select Target Dataset” panel and select the “RNAseq” data type.



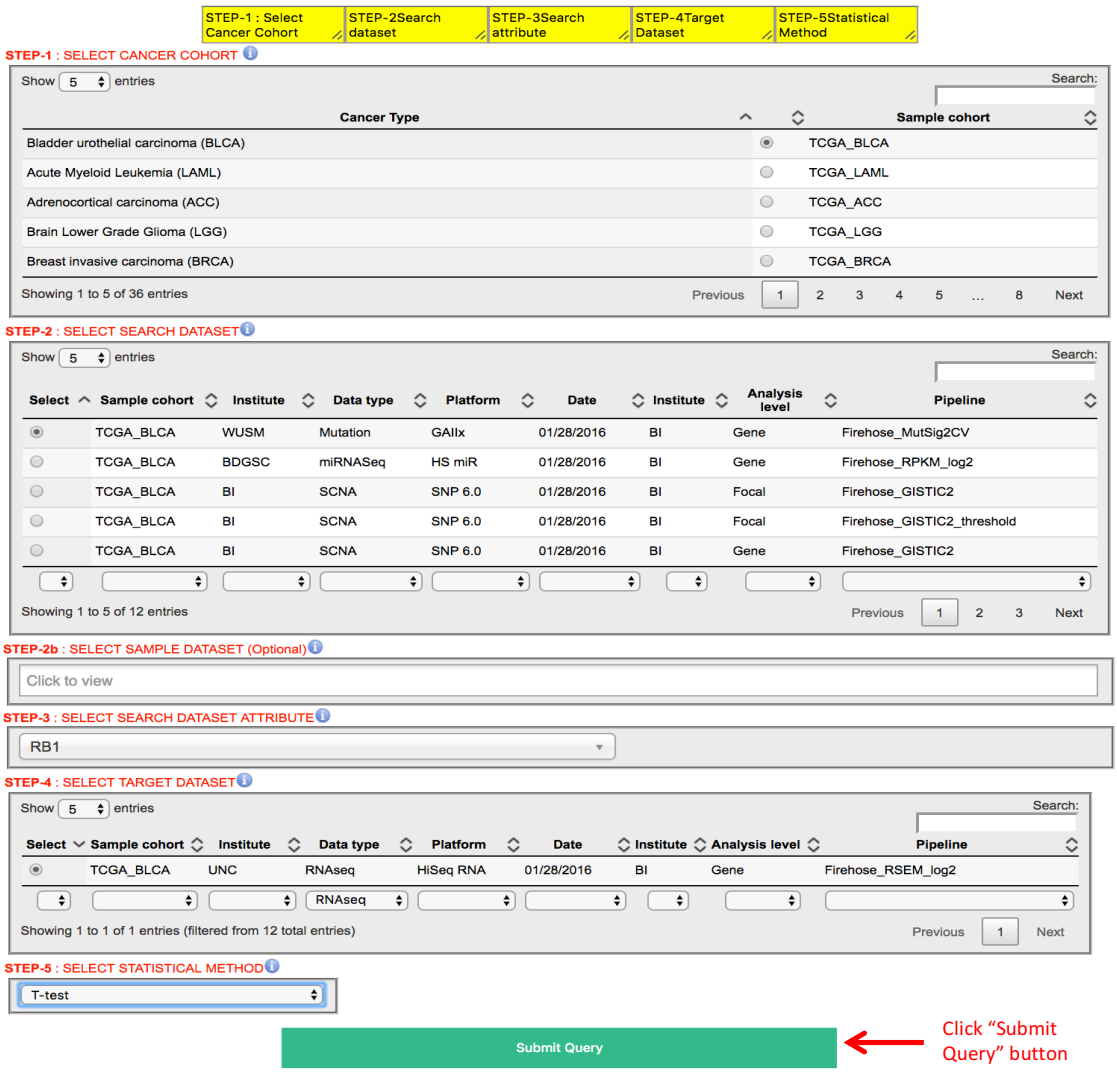
6. Select the gene level “RNAseq” dataset, which was obtained from the Broad Firehose with release date “01/28/2016”.



7. After selecting the target dataset, the “Select Statistical Method” panel will appear. Select the appropriate statistical method from the dropdown menu. T-Test is selected from the menu panel.

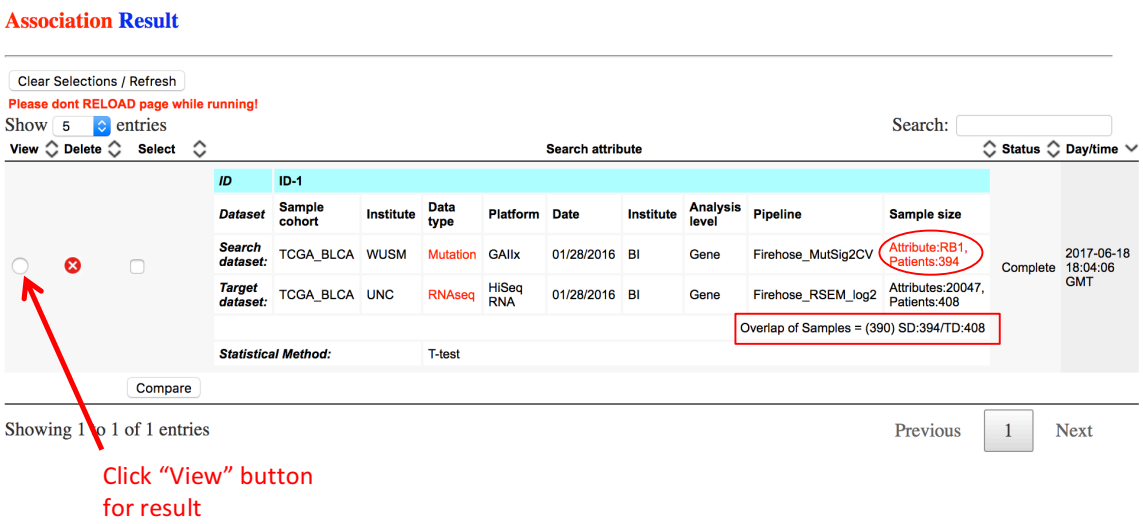


8. Click “Submit Query”.

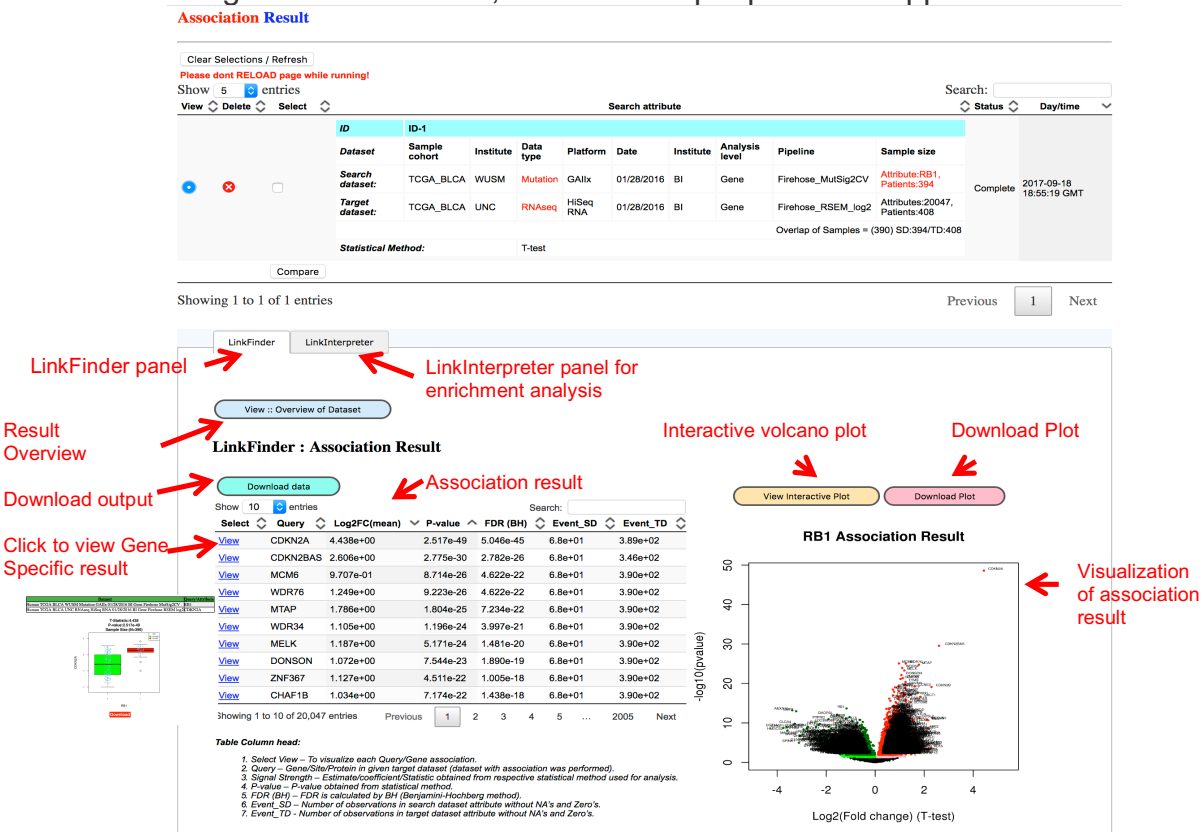


9. On Submit, the page will direct to the result page, where the analysis is performed on-the-fly. Do not reload the page while the program is running. The

result panel has the (i) search dataset, (ii) target dataset, (iii) selected attribute, (iv) samples in each dataset, (v) overlapping samples in both datasets, (vi) the selected statistical method, and a (vii) timestamp (GMT). To view the result, click on the “View” radio button, which is on the left side in the “View” column.



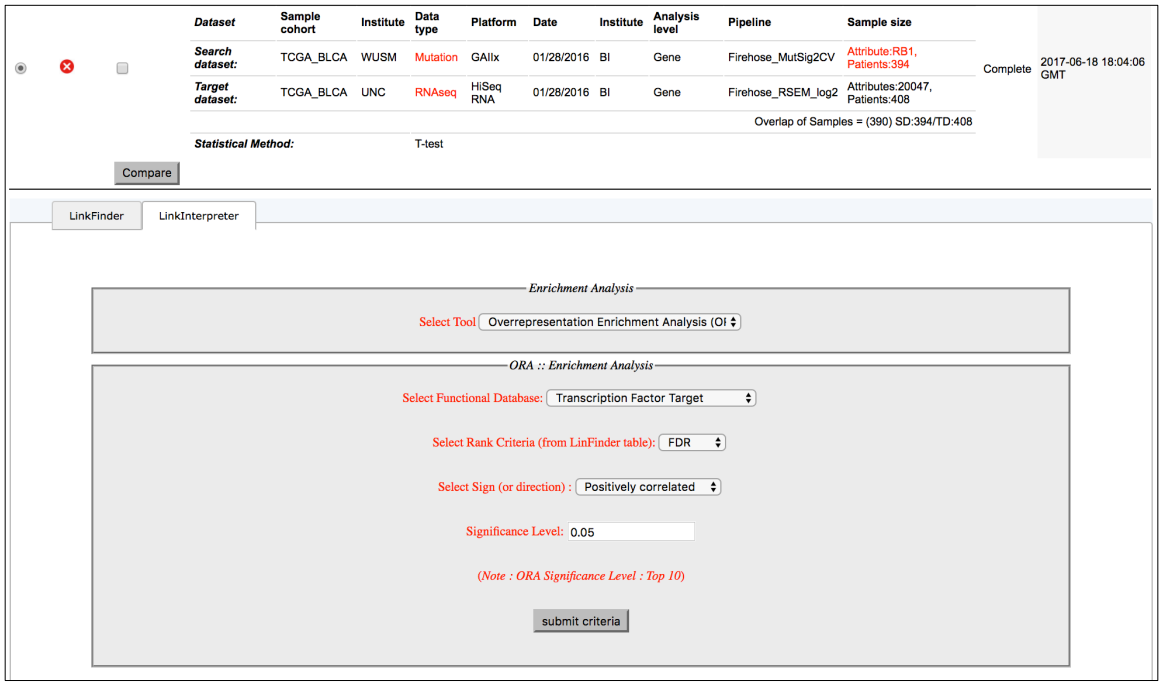
10. After selecting the “View” button, the result output panel will appear.



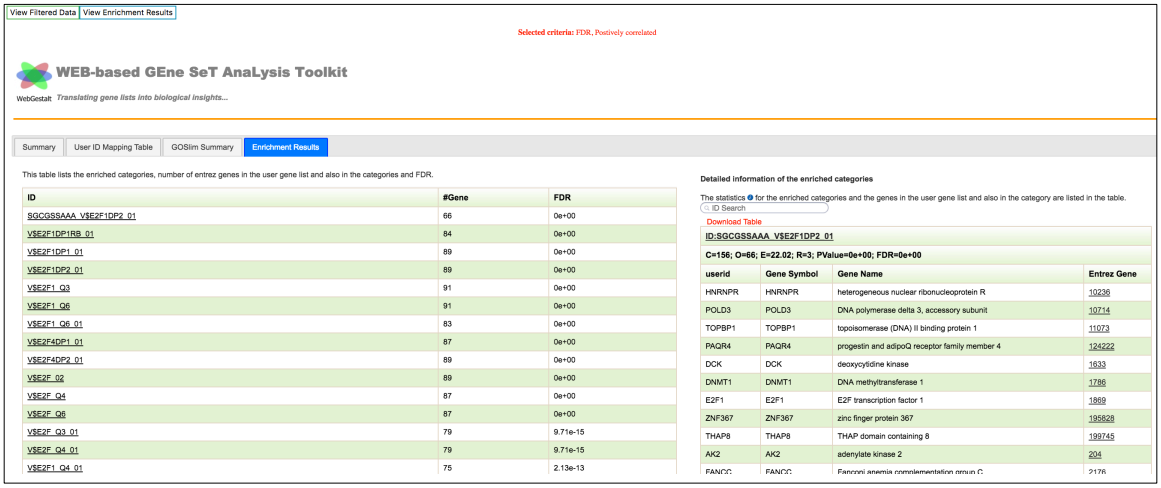
11. Output: The result is sorted based on the Statistic (descending) and P-value (ascending) columns.

1. Select View – To visualize each Query/Gene association.
2. Query – Gene/Site/Protein in given target dataset (dataset with association was performed).
3. Signal Strength – Estimate/coefficient/Statistic obtained from respective statistical method used for analysis.
4. P-value – P-value obtained from statistical method.
5. FDR (BH) – FDR is calculated by BH (Benjamini-Hochberg method).
6. Event_SD – Number of observations in search dataset attribute without NA's and Zero's.
7. Event_TD - Number of observations in target dataset attribute without NA's and Zero's.

12. For enrichment analysis select "Overrepresentation analysis" method and select "Transcription factor target" as a functional database. Select "FDR" as rank criteria for "positively correlated genes". Select significance level 0.05.



13. Output results in enrichment results showing transcriptional factor "E2F1" enriched in positively correlated genes as targets.



14. Correlation between *RB1* mutation and mRNA expression in BRCA (*N*=972) and BLCA (*N*=390) is performed. We combined the expression signatures from *RB1* mutated samples found in BRCA and BLCA together to increase the power of identifying common targets in both of the cancer cohort.


(A)
BLCA Cohort
RB1 Mutation vs. mRNA Expression
BRCA Cohort
RB1 Mutation vs. mRNA Expression
-log10(pvalue)
Log2(Fold change) (T-test)
(B)
1
2
3
Compare
ID
ID-2
Dataset
Sample cohort
Institute
Data type
Platform
Date
Institute
Analysis level
Pipeline
Sample size
Search dataset:
TCGA_BRCA
WUSM
Mutation
GAIIx
01/28/2016
BI
Gene
Firehose_MutSig2CV
Attribute:RB1, Patients:975
Target dataset:
TCGA_BRCA
UNC
RNAseq
HiSeq RNA
01/28/2016
BI
Gene
Firehose_RSEM_log2
Attributes:20155, Patients:1093
Overlap of Samples = (972) SD:975/TD:1093
Statistical Method:
T-test
Complete
2017-06-18 18:05:19 GMT
ID
ID-1
Dataset
Sample cohort
Institute
Data type
Platform
Date
Institute
Analysis level
Pipeline
Sample size
Search dataset:
TCGA_BLCA
WUSM
Mutation
GAIIx
01/28/2016
BI
Gene
Firehose_MutSig2CV
Attribute:RB1, Patients:394
Target dataset:
TCGA_BLCA
UNC
RNAseq
HiSeq RNA
01/28/2016
BI
Gene
Firehose_RSEM_log2
Attributes:20047, Patients:408
Overlap of Samples = (390) SD:394/TD:408
Statistical Method:
T-test
Complete
2017-06-18 18:04:06 GMT
Showing 1 to 3 of 3 entries (filtered from 57 total entries)
Previous
1
Next
(C)
LinkCompare
LinkInterpreter
K
LinkCompare :: Comp
I
e associations
J
D
Table
Scatter Plot
Venn Plot
Download Table
H
F
Show 10 entries
Search:
E
G
Query (Gene/Protein/site)
ID-2-SignalStrength
ID-2-Pvalue
ID-2-FDR
ID-1-SignalStrength
ID-1-Pvalue
ID-1-FDR
Sumz_stat
Sumz_P
Sumz_FDR
CDKN2A 9.687e-01 4.559e-02 3.501e-01 4.438e+00 2.517e-49 5.046e-45 1.185e+01 2.075e-32 4.151e-28
MCM6 8.425e-01 1.438e-05 1.187e-03 9.707e-01 8.714e-26 4.622e-22 1.049e+01 9.454e-26 9.456e-22
WDR76 8.144e-01 2.659e-05 1.983e-03 1.249e+00 9.223e-26 4.622e-22 1.039e+01 2.732e-25 1.822e-21
DONSON 1.084e+00 2.449e-06 2.771e-04 1.072e+00 7.544e-23 1.890e-19 1.029e+01 7.794e-25 3.898e-21
MELK 1.226e+00 1.781e-05 1.441e-03 1.187e+00 5.171e-24 1.481e-20 1.018e+01 2.428e-24 9.713e-21
CHAF1B 8.333e-01 3.538e-06 3.711e-04 1.034e+00 7.174e-22 1.438e-18 1.007e+01 7.137e-24 2.379e-20
MCM2 9.677e-01 1.166e-06 1.535e-04 9.832e-01 5.717e-20 7.969e-17 9.907e+00 3.875e-23 1.107e-19
CDKN2BAS 1.021e+00 1.520e-02 1.969e-01 2.606e+00 2.775e-30 2.782e-26 9.803e+00 1.095e-22 2.433e-19
TYMS 1.033e+00 8.570e-06 7.739e-04 1.240e+00 4.139e-21 7.544e-18 9.814e+00 9.776e-23 2.433e-19
E2F1 1.066e+00 1.027e-06 1.389e-04 1.160e+00 1.686e-18 1.536e-15 9.661e+00 4.411e-22 8.823e-19
Showing 1 to 10 of 20,003 entries
Previous 1 2 3 4 5 ... 2001 Next
Note:
ID-XX-SignalStrength: Signal strength (like Log2FC, log(Odd ratio), F-Statistic, ...) obtained from statistical method for given ID-xx (Refer to individual ID for query information).
ID-XX-Pvalue: P-value obtained from statistical method for given ID-xx.
ID-XX-FDR: FDR is calculated by BH (Benjamini-Hochberg method) for given ID-xx.
Sumz_stat: Stouffer method (Stouffer, 1949) based meta-analysis estimate
Sumz_P: P-value obtained from Stouffer method (Stouffer, 1949) based meta-analysis

5. Data Sharing and Annotation

This document describes the requirements for submission of omics data from new cohorts into the LinkedOmics database. The goal is to have a reasonably unified set of labels for the files with associated metadata for easy accessibility to users. Please send a request for data integration to linkedomics.zhanglab@gmail.com, before transferring data. The shared data can remain private or be opened for public use based on the sender's instructions.

5.1 Data Format

Required files are the (i) annotation file, and (ii) data file. The data file should be submitted in matrix format in excel or csv file format. The genes or attributes should be in rows and sample names should be in columns. Each data type should be in a different file. For example for breast cancer gene level RNAseq data,

Folder: Breast Cancer
|- Mutation
|--- Gene Level
|--- Site Level

5.2 Data Submission Guidelines

The annotation file can be created in text (.txt) or excel (.xls). The file should have 10 columns. The 10 columns (categories) are discussed below.

No.	Category	Description
(i)	ID	Specific assigned ID
(ii)	Species	Organism like *Homo sapiens*, *Mus musculus*
(iii)	Sample Cohort	Cohort from which samples are obtained (related to the cancer study)
(iv)	Institute(Laboratory)	Associated with a specific laboratory where the experiments were run
(v)	Data Type	Omics data type (e.g. mRNA, proteome, phosphoproteome)
(vi)	Platform	Platform on which experiments were carried out (e.g. IlluminaHiSeq, HumanMethylation27)
(vii)	Date/Version	Date of accession of data type
(viii)	Analysis Institute	Associated with the specific institute that processed the data
(ix)	Analysis Level	Data analysis at gene, site, isoform, analyte, or focal level
(x)	Pipeline	Version of the data type or specific method used for analysis (or extra comments can be added)

File Names

The proposed file names are composed of 10 categories, with an underscore separating each category. For example user wants to submit methylation27 data at CpG site level into the database such as,

01 Human_TCGA BRCA JHU/USC Methylation Meth27 01/28/2016 BI CpG Firehose_Methylation_Preprocessor_v1

Lets look at each category,

i. [**01, 02, 03…**]
   Two-digit number that represents the ID number of the given data type/file name.
ii. **Human** labels the species name.
iii. **TCGA_BRCA** represents the cancer cohort and its source. Here, the Breast Cancer cohort from TCGA.
iv. **JHU/USC** labels the center that performed the experiments for a given data type (JHU-Johns Hopkins University, USC-University of South Carolina).
v. **Methylation** is the data type representing the methylation study on cancer patient samples.
vi. **Meth27** labels the platform of the experimental run (here IIIlumina Infinium Human DNA Methylation27 platform).
vii. **01/28/2016** labels the date a given dataset was procured in the format MM/DD/YYYY.
viii. **BI** labels the institute that analyzed the dataset (here the Broad Institute). (Others can be JHU-Johns Hopkins University; PNNL-Pacific Northwest National Laboratory; UNC-University of North Carolina; WU-Washington University in St. Louis)
ix. **CpG** labels the level at which the data are analyzed (here CpG site level).
x. **Firehose_Methylation_Preprocessor** labels the pipeline or different processing version for a given data type (here methylation preprocessor filters methylation data for use in downstream pipelines). Optional version numbers, like **v1,** designates the file version by a given center. Or user can submit preferred naming foe given datatype.

References

- Broad Institute (http://gdac.broadinstitute.org/).
- TCGA data portal (http://cga-data.nci.nih.gov/tcga).
- CPTAC data portal (https://cptac-data-portal.georgetown.edu/cptacPublic/).
- WebGestalt (http://www.webgestalt.org)
- Benjamini and Hochberg 1995
- Stouffer S, DeVinney L, Suchmen E. The American soldier: Adjustment during army life. Vol. 1. Princeton University Press; Princeton, US: 1949.
- Dewey M. metap: meta-analysis of significance values. R package version 0.8 2017.