

# NetGestalt User Manual

July 8, 2016

## Table of Contents

I. Introduction .....	3
II. User Interface .....	3
1. Select a portal .....	3
2. Set a view .....	4
a. Select a view .....	4
b. Upload a view .....	5
c. Delete a view .....	6
3. Add tracks.....	6
a. Browse system tracks.....	6
b. Search tracks .....	7
c. Upload tracks.....	7
(i) Composite continuous track (CCT) file.....	8
(ii) Composite binary track (CBT) file .....	9
(iii) Track sample information (TSI) file .....	9
(iv) Single continuous track (SCT) file.....	10
(v) Single binary track (SBT) file.....	10
d. Enter gene symbols as a track .....	11
4. Visualize Tracks .....	12
a. CCTs.....	12
b. CBTs.....	12
c. SCTs .....	12
d. SBTs .....	12
5. Zoom in (out) tracks .....	13
a. Click bars representing predefined modules.....	14
b. Alt+drag.....	14
c. Double click.....	14
d. Pan.....	14
6. Resize tracks .....	15

7.	Hierarchically cluster tracks.....	16
a.	Clustering w/respect to system views .....	16
b.	Clustering w/respect to user views .....	16
8.	Analyze tracks .....	17
a.	Network analysis .....	17
(i)	Module enrichment .....	17
(ii)	Network expansion.....	19
(iii)	Gene prioritization .....	19
b.	Gene Set Enrichment.....	19
c.	Subtrack annotation .....	20
d.	Statistical analysis .....	20
e.	Data transformation .....	20
f.	Value-based filtering.....	21
g.	Presence-based filtering.....	22
h.	Node-link Graph .....	22
i.	Zoom to a gene .....	23
j.	Track comparison .....	24
k.	Track co-visualization.....	25
l.	Switch between different views.....	25
III.	Portal descriptions.....	27
1.	Generating Views .....	27
2.	The Colorectal Cancer (CRC) portal track description .....	27
a.	Genomic and proteomic alterations in the TCGA CRC tumor cohort.....	28
b.	Clinical relevance based on CRC tumor tissue .....	30
c.	Tracks based on CRC cell lines.....	32
d.	Functional tracks .....	32
3.	The Cancer Genome Atlas (TCGA) Portal.....	33
a.	TCGA profile tracks .....	34
b.	TCGA clustering tracks.....	35
c.	TCGA portal TSI track data sources .....	36
d.	Statistical correlation results.....	36
4.	Clinical Proteomic Tumor Analysis Consortium (CPTAC) Portal .....	37
a.	Protomic, phosphoproteomic, and glycoproteomic alterations from CPTAC cohorts.....	38

i. Colorectal proteomic profile tracks.....	39
ii. Ovarian and Breast Proteomic profile tracks .....	40
iii. TCGA profile tracks in CPTAC portal.....	41
b. CPTAC portal tsi track data sources .....	43
c. Statistical correlation results.....	45
References:.....	47

## I. Introduction

NetGestalt is a novel data integration framework that allows simultaneous presentation of large scale experimental and annotation data from many sources in the context of biological networks or genomes to facilitate data interpretation and hypothesis generation.

The NetGestalt framework provides various features for data query/upload, visualization, and integration. This manual introduces all features that can be accessed through the user interface (Section II) as well as several portals developed based on the NetGestalt framework (Section III).

## II. User Interface

### 1. Select a portal

There are multiple portals available for NetGestalt. Each portal contains both the protein-protein interaction network “views” for a given species along with a chromosome “view”, as well as a large set of functional information from sources such as KEGG, Gene Ontology, and Drugbank. Currently, there are separate portals for the following species: *Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Danio rerio*.

In addition, there are two human portals populated with publicly available data: The “Human: Colorectal Cancer Portal”, which includes omics data from various colorectal tumor cohorts including The Cancer Genome Atlas (TCGA) cohort as well as colorectal cancer cell lines, and the “Human: CPTAC Portal”, which includes data from both the TCGA study and the Clinical Proteomics Tumor Analysis Consortium (CPTAC) study of human breast, colon, and ovarian cancer.

Users must select a portal on the Netgestalt home page (Figure 1).

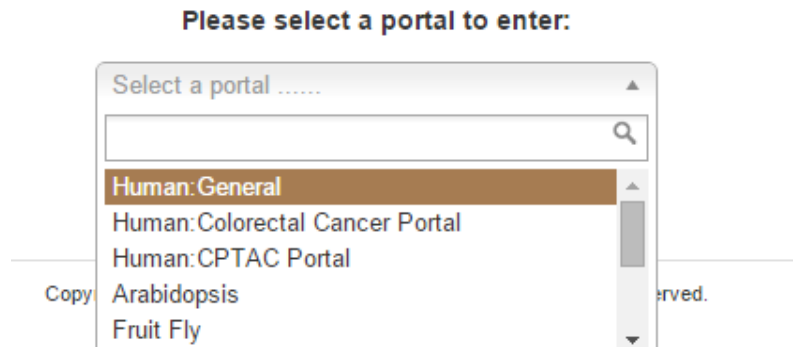


Figure 1. Screen image of portal selection dropdown box.

## 2. Set a view

To perform analysis in NetGestalt, a user should first select a view. When hovering the mouse over the “View” menu (see red box in Figure 2), different choices for setting the view will be shown in a drop-down menu.

### a. Select a view



Figure 2. Set a view in NetGestalt.

When hovering the mouse over “Select”, all views provided by the system will be shown in a menu. The currently active view is shown in grey, while others are shown in black.

The user can select a view by clicking the name. After setting the view, the user can find the name of the selected view and corresponding category at the top-right of the window (see brown box in Figure 2). In the current version, NetGestalt contains two categories of views: network views and chromosome views. Each portal contains a single chromosome view and at least one network view. For the human portals, the “hprd” and “iRef”, correspond to the HPRD human protein-protein interaction (PPI) network (<http://www.hprd.org/>) and iRef human PPI network (<http://wodaklab.org/iRefWeb/>), respectively.



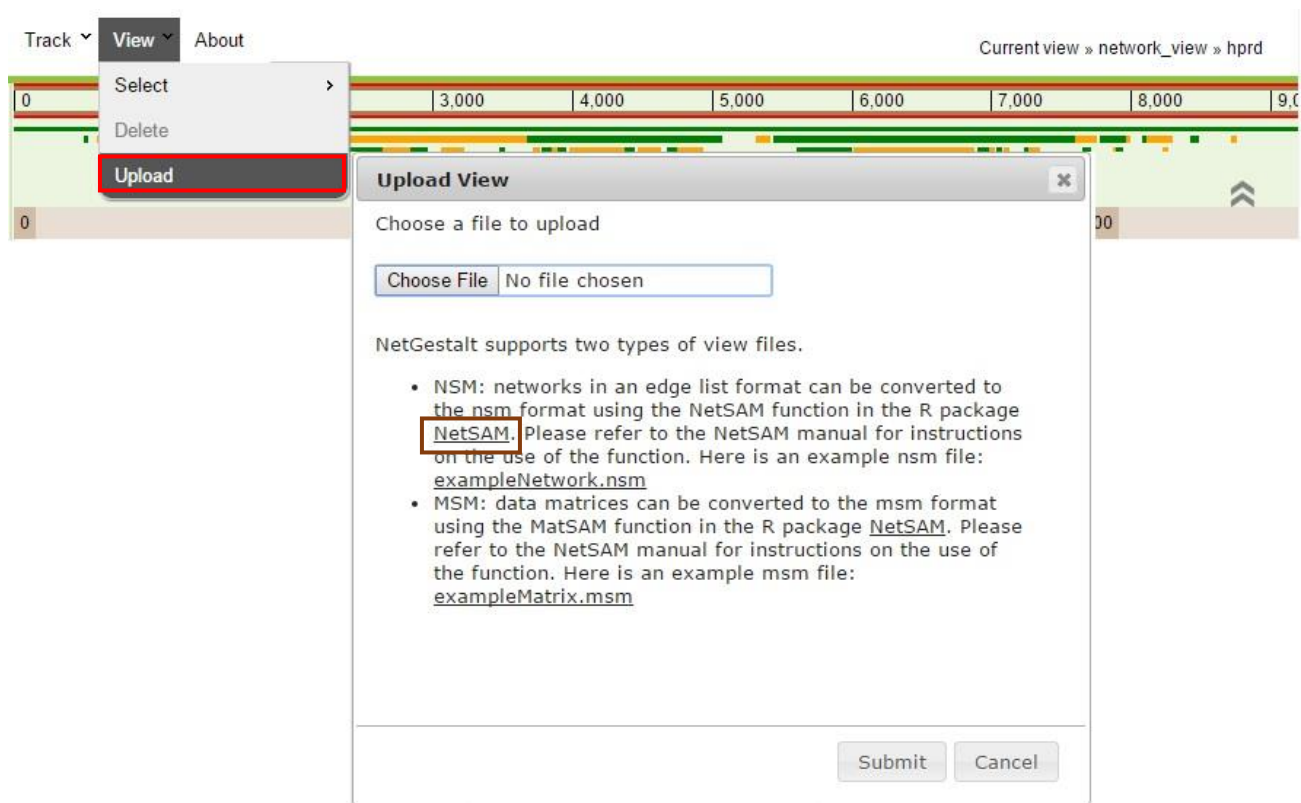
**Figure 3. Examples of a “network view” (top), derived from the iREF protein-protein interaction network, and a “chromosome view” (bottom).**

After selecting a “network\_view”, a one-dimensional layout of network nodes will be shown right below the menu (Figure 3, the ruler at the top of the upper panel). Below the ruler, bars in different length and thickness represent modules at different hierarchical levels of the network. The thick bars correspond to modules in the best partition of the global network. Modules at this level can be split into smaller ones represented by thin bars. Alternating bar colors (green and orange) are used to help users distinguish neighboring modules. If a “chromosome\_view” is selected (Figure 3, lower panel), the thick bars correspond to chromosomes in numbered order from left to right. The smaller bars correspond to chromosomal bands. The module information can be hidden clicking the grey “double arrow” button at the right bottom of this region (see blue boxes in Figure 3).

#### **b. Upload a view**

Users can also upload their own views into NetGestalt by clicking the “Upload” button below the “Select” button in the menu (see red box in Figure 4). After clicking, an “Upload View” dialog window will pop up. In this window, a user can click the “Choose File” button to select a local file and then click the “Submit” button to upload the view to NetGestalt. **NetGestalt accepts two types of view files: “.nsm” network files or “.msm” data matrix files input (network and track information combined into a single file), which can both be generated by the R package NetSAM** (<http://www.bioconductor.org/packages/release/bioc/html/NetSAM.html>).

The NetSAM package as well as the NetSAM manual can be downloaded in the “Upload View” dialog window (see brown box in Figure 4). After uploading the network, the current view will be automatically switched to the uploaded network view, which can also be found in the drop down menu under “View” → “Select”.



**Figure 4. Upload a “view” in NetGestalt.**

### c. Delete a view

Only user-uploaded views can be deleted. To delete an uploaded network view, the user has to switch to another network view first and then click the “delete” button below the “Select” button in the menu under “View” and select the view they would like to delete. If you are deleting a view currently in use you will be asked to select a remaining view to switch to.

## 3. Add tracks

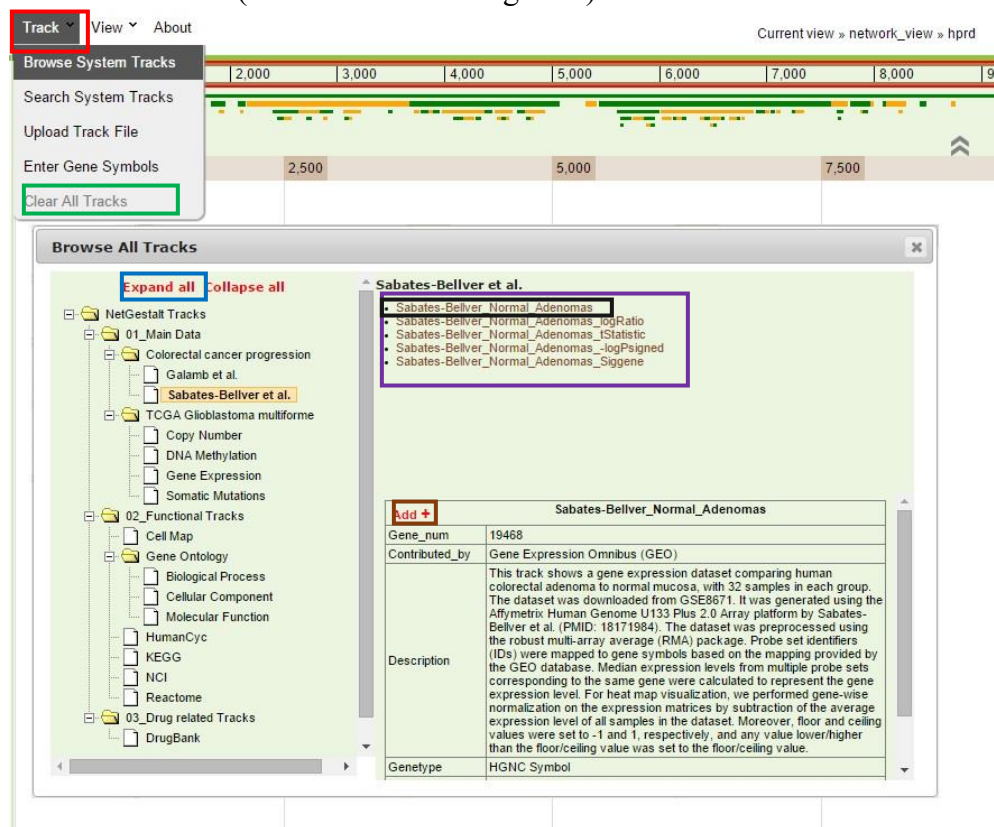
NetGestalt provides multiple options for adding tracks to the track viewing area. Hovering the mouse over the “Track” menu (see red box in Figure 5), different options for retrieving and adding tracks of interest will be shown in a drop-down menu.

### a. Browse system tracks

A user can click the “Browse System Tracks” button to browse all the tracks included in the NetGestalt database. In the “Browse All Tracks” dialog, all the tracks are organized into a tree structure.

The user can click the “Expand all” button (see blue box in Figure 5) to expand the tree and select a category of tracks by clicking a leaf node of the tree (such as the leaf node “Sabates-Bellver et al.”). Then, a list of tracks contained in the category will be shown in the top-right part of the dialog (see purple box in Figure 5). When clicking one track in the list, such as

the “Sabates-Bellver\_Normal\_Adenomas” track (see black box in Figure 5), detailed information associated with the track will be shown in the bottom-right table. Finally, clicking the “+” button. (see brown box in Figure 5) will add the track to the track viewing area.



**Figure 5. Browse tracks in NetGestalt.**

## b. Search tracks

A user can click the “Search System Tracks” button below the “Browse System Tracks” button to search for tracks of interest in the database. After clicking the “Search System Tracks” button, a “Search system tracks” dialog will open. The user can input a key word in the box (see red box in Figure 6). NetGestalt will list the names of all matching tracks below the box. Hovering the mouse over the “i” button of a track will display the detailed information about the track. Finally, the track can be added to the track viewing area by clicking the “+” button.

## c. Upload tracks

Users can upload their own data into NetGestalt by clicking the “Upload Track File” button below the “Search System Tracks” button (see red box in Figure 7). In the “Upload Track” dialog, the user can click the “Choose File” button to select a local file (see blue box in Figure 7), then select the type of gene identifier from the “ID type” drop down box (see orange box in Figure 7), and then finally click the “Submit” button to upload the file to NetGestalt. For network analyses based on user-uploaded networks, the user can use any type of IDs matched with the IDs in the user-uploaded data tracks. NetGestalt supports five types of track files corresponding to different data types. Please see below for file

preparation guidelines. Examples for the five types of track files can be downloaded from the “Upload Track” dialog (see brown box in Figure 7).

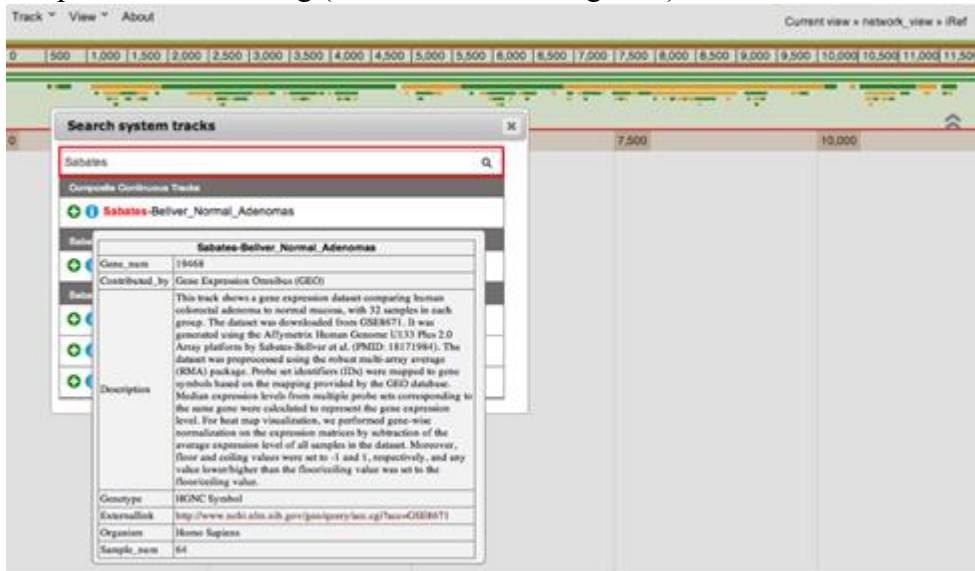


Figure 6. Search tracks in NetGestalt.

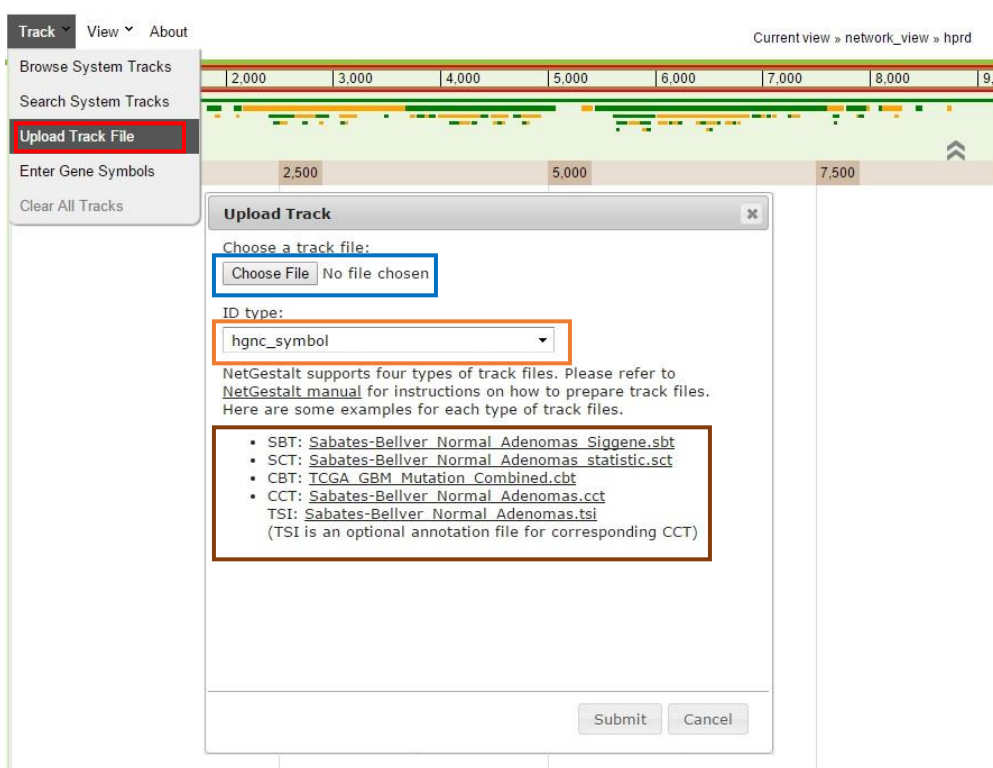


Figure 7. Upload tracks in NetGestalt.

### (i) Composite continuous track (CCT) file

**Definition:** a CCT file (.cct) is a tab-delimited text file that contains data for a composite track with multiple related sub-tracks with continuous data (e.g. microarray gene expression



data for samples from the same data set). The file name will be used as the track name.

File format description: a CCT file is a data matrix in which each row represents a gene and each column represents a sample. The first column lists the gene ids (e.g. gene symbols) and the first row lists the sample names. Two or more data columns (sub-tracks) are required. Columns must be separated by tab. Each cell in the matrix is a continuous value for corresponding gene and sample. Missing values are represented by NA. Data for different samples must be comparable (i.e. properly normalized). Duplicated row names or column names are not allowed. No special characters for row or column names.

Example:

GeneSymbol	Sample1	Sample2	Sample3	Sample4
Gene1	0.025	-0.55	-1	0.095
Gene2	-0.077	0.069	0.64	0.18
Gene3	-0.47	1	0.87	-0.88
Gene4	-0.71	-0.19	0.33	-0.45

### **(ii) Composite binary track (CBT) file**

Definition: a CBT file (.cbt) is a tab-delimited text file that contains data for a composite track with multiple related sub-tracks with binary data (e.g. mutation status for genes in multiple samples). The file name will be used as the track name.

File format description: same as the CCT file except that each cell in the matrix is a binary value (0 or 1) for corresponding gene and sample.

### **(iii) Track sample information (TSI) file**

Definition: a TSI file (.tsi) is a tab-delimited text file that contains the sample information for a CCT or CBT. This file is an optional sample annotation file for the matching CCT/CBT file, and it should be uploaded together with the CCT/CBT file.

File format description: a TSI file (.tsi) is a data matrix in which each row represents a sample and each column represents one feature of the samples. Sample features can be divided into five data types: binary data (BIN, e.g., mutation status), categorical data (CAT, e.g., tumor stage), continuous data (CON, e.g., age), survival data (SUR, e.g. overall survival), and paired binary data (PAIRED, e.g. paired normal vs tumor). Binary data do not have to be 0/1 but must contain exactly two categories (e.g. yes/no or tumor/normal). The first column lists the sample names. Sample names must match exactly those in the corresponding CCT or CBT. The first row lists the feature names; and the second row indicates the data type for each feature (must be one of the following: BIN, CAT, CON, SUR, or PAIRED). Each cell in the matrix is a value for corresponding sample and feature. For survival data, time and event are separated by “,”. For paired binary data, the pair id and binary value are separate by “,”. Missing values are represented by NA. Duplicated row names or column names are not allowed. No special characters for row or column names.

Example:

Barcode	Age	Anatomic neoplasm	Colonpolyps	Overall survival	Tissue
data_type	CON	CAT	BIN	SUR	PAIRED
TCGA-A6-2670-01	45	sigmoid colon	0	259,0	1,batch1
TCGA-A6-2671-01	85	sigmoid colon	0	437,0	1,batch2
TCGA-A6-2672-01	82	transverse colon	0	1321,1	2,batch1
TCGA-A6-2674-01	71	sigmoid colon	0	NA,NA	2,batch2
TCGA-A6-2675-01	78	sigmoid colon	0	434,0	3,batch1
TCGA-A6-2676-01	75	cecum	0	1437,1	3,batch2
TCGA-A6-2677-01	68	cecum	0	635,0	4,batch1
TCGA-A6-2678-01	43	transverse colon	0	437,0	4,batch2
TCGA-A6-2679-01	73	ascending colon	0	145,0	5,batch1
TCGA-A6-2680-01	72	hepatic flexure	0	465,1	5,batch2
TCGA-A6-2681-01	73	cecum	0	882,1	6,batch1
TCGA-A6-2682-01	74	cecum	0	889,1	6,batch2

**(iv) Single continuous track (SCT) file**

Definition: a SCT file (.sct) is a tab-delimited text file that contains statistical analysis results derived from a data set (e.g. mean, median, sum, variance, t-statistic, p value).

File format description: an SCT file contains at least two columns: the first column with gene ids (e.g. gene symbols) and the second column with statistic scores (data) for the genes. Up to three statistic scores (three data columns) can be included in an SCT file. Columns must be separated by tab. The first row of the file lists the track names. We suggest that a track name should contain some descriptions about the corresponding statistic score. For fold changes, it is recommended to perform a log<sub>2</sub> transformation. For p values, it is recommended to perform a -log<sub>10</sub> transformation and add the direction of change to the transformed p values (e.g. signed p values). Missing values are represented by NA. Duplicated row names or column names are not allowed. No special characters for row or column names.

Example:

GeneSymbol	TrackName1	TrackName2
Gene1	-2.2	-4.8
Gene2	-1.3	-1.5
Gene3	5.3	7.6
Gene4	1.1	0.7

**(v) Single binary track (SBT) file**

Definition: a SBT file (.sbt) is a tab-delimited text file that contains lists of genes in separate rows (e.g. significant genes from differential expression analysis).

File format description: an SBT file contains track name, track description and gene ids (e.g.

gene symbols) in the track. Each row represents a track and columns are separated by tab. Up to three tracks (three rows) can be included in an SBT file. To enable meaningful enrichment analysis, the user can include in the first row an “All” track that contains the reference gene symbols for the tracks in the SBT file (e.g. all genes on the microarray platform from which the differentially expressed genes were identified). If this information is not provided, enrichment analysis will be based on all genes in the network. If the “All” track is provided, all genes in the other tracks should be included in the “All” track. No special characters for Track names. Cells in red should not be changed.

Example:

All	Description	GeneSymbol_11	GeneSymbol_21	GeneSymbol_22
TrackName1	Description1	GeneSymbol_11		
TrackName2	Description2	GeneSymbol_21	GeneSymbol_22	

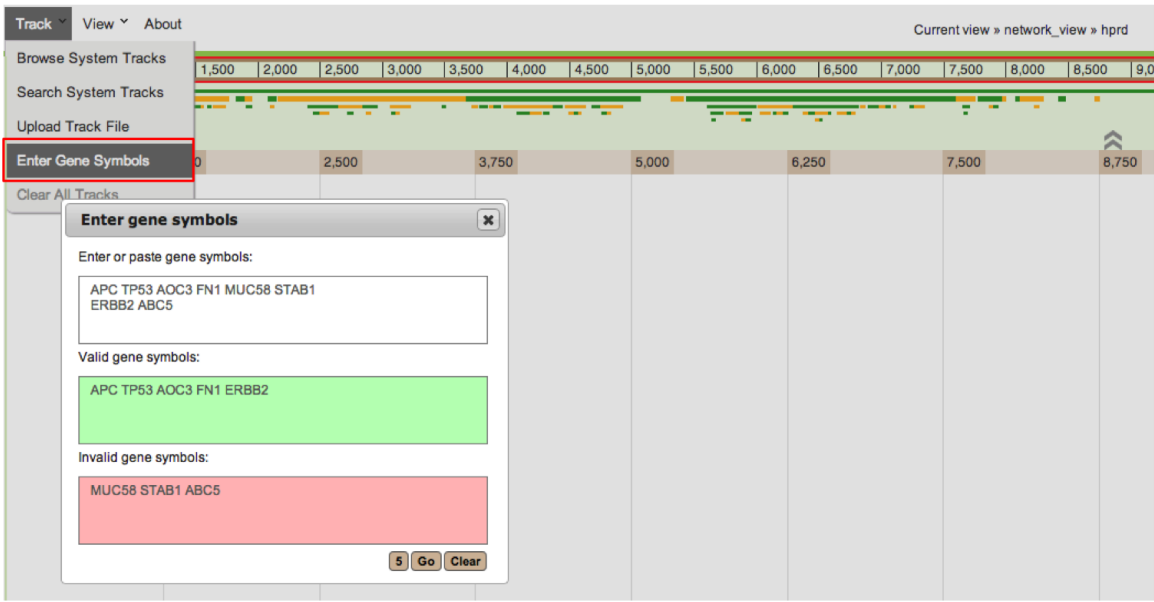
NetGestalt uses the following rules to determine the file type of a given file:

- (1) Use the file extension to determine the file type. Ignore the .txt file extension. (For example, both test.sbt and test.sbt.txt are treated as a SBT file.)
- (2) If that fails, NetGestalt cannot determine the file type and displays an error message.

The maximum upload file size is 50MB.

**d. Enter gene symbols as a track**

To enter gene symbols and add them as a new SBT in NetGestalt, the users can click the “Enter gene symbols” button in the “Track” menu (see red box in Figure 8) and then enter or paste gene symbols in the “Enter gene symbols” dialog. The input will be automatically separated into valid gene symbols (i.e. gene symbols included in the current view) and invalid ones. The total number of valid gene symbols will be shown at the bottom of the window. Clicking the “GO” button, entering a track title in the opened dialog, and then clicking the “Add” button will add the valid gene symbols as a new SBT (Figure 8).



**Figure 8. Add gene symbols as a track in NetGestalt.**

## 4. Visualize Tracks

NetGestalt uses different methods to visualize different types of tracks.

### a. CCTs

NetGestalt visualizes a CCT track with a heat map with colors ranging from blue to red. The first track in Figure 9 (green box) is a CCT track representing a gene expression data containing 32 CRC samples and 32 normal mucosa samples. When hovering the mouse over the heat map plot, NetGestalt will show the gene id, sample index and sample name at the corresponding position.

### b. CBTs

NetGestalt visualizes a CBT track with a heat map of two colors (red and grey). The fourth track in Figure 9 (yellow box) shows a CBT track representing a TCGA GBM somatic mutation data containing 148 CRC samples. Hovering the mouse over the heat map plot will show the gene symbol, sample index and sample name at the corresponding position.

### c. SCTs

NetGestalt visualizes a SCT track with a bar plot. The second track in Figure 9 (purple box) is an SCT track containing  $-\log(p \text{ value})$  of gene expression between CRC samples and normal samples based on Sabates-Bellver dataset (GSE8671). Hovering the mouse over the bar plot, NetGestalt will display the gene symbol and statistic value at the corresponding position.

### d. SBTs

NetGestalt visualizes a SBT track with a barcode plot. The third track in Figure 9 (black box) is an SBT track representing significantly genes based on Sabates-Bellver dataset. Hovering

the mouse over the plot will display gene symbol at the corresponding position.



**Figure 9. Visualize four types of tracks in NetGestalt. The top track is an example of a Composite Continuous Track (CCT) showing continuous data per gene (x-axis) per sample (y-axis) (green box). Below is a Single Composite Track (SCT) showing one continuous value (y-axis) per gene (axis) (purple box). Next, a Single Binary Track (SBT) shows binary data per gene via the presence or absence of a vertical black (black box). Finally, a Composite Binary Track (CBT) shows binary data per gene (x-axis) per sample (y-axis), where a “true” value is presented by a red dot and “false” is gray (yellow box).**

After adding the tracks into the track viewing area, NetGestalt automatically shows the track name on the top-left of the track (see red box in Figure 9). User can click the “double arrow” button located above the top-left of the first track to hide these track names (see blue box in Figure 9). Dragging a track name can change the vertical position of the track (not supported by IE).

Hovering the mouse over a track name, all track manipulation and analysis features will be shown in a drop-down menu (see orange box in Figure 9).

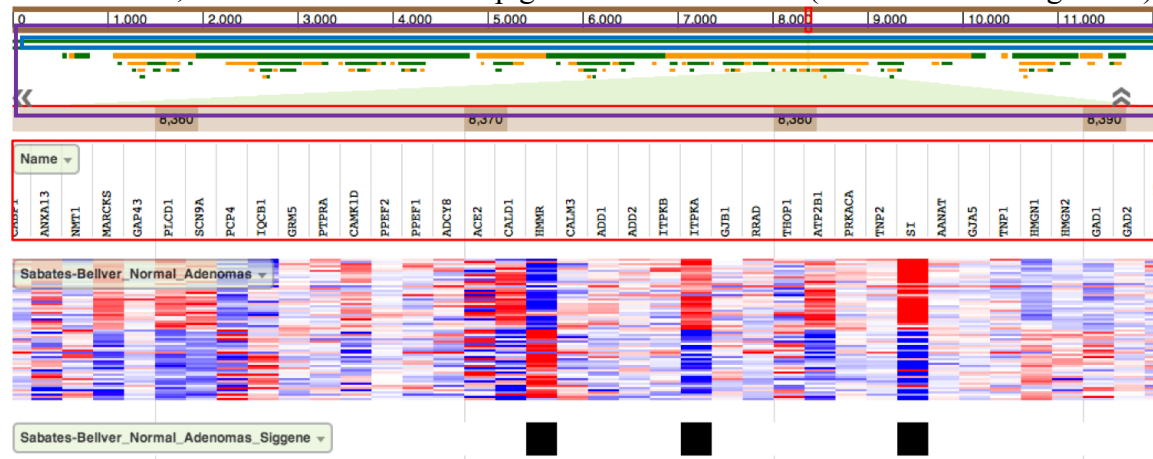
When clicking the “i” button (see brown box in Figure 9), a table containing detailed information about the track will be shown. Clicking the “e” button will export the track data (see gray box in Figure 9). Clicking the “x” button will remove the corresponding track from the viewing area (see gold box in Figure 9). We will introduce other buttons in the section “Analyze tracks”. The user can also click the “Clear all tracks” button in the “Track” menu to remove all the tracks from the track viewing area (see green box in Figure 5).

## 5. Zoom in (out) tracks

NetGestalt provides multiple methods to visualize tracks at different scales described below.

### a. Click bars representing predefined modules

NetGestalt uses horizontal bars to represent network modules (sub-networks) at different hierarchical levels (see purple box in Figure 10). The lengths of the bars correspond to the size of the modules. For a specific hierarchical level, genes within a module are highly connected whereas genes from different modules are loosely connected. Most of the predefined network modules have been demonstrated to be functionally, spatially or dynamically homogeneous (Shi et al., Nat Methods 2013). These modules can help users to easily associate subnetworks with experimental data. User can click the corresponding bar to zoom into a module for further analysis (see Figure 10). A track of gene symbols will appear at the top of the track viewing area when it is completely zoomed in (see red box in Figure 10). To zoom out, the user can click the top green bar at root level (see blue box in Figure 10).



**Figure 10.** Click bars representing predefined modules to zoom in the tracks.

If a user is interested in a region that is not represented by a predefined module, NetGestalt provides two additional zoom-in methods for visualizing any regions of the one-dimensionally ordered network (See below).

### b. Alt+drag

A user can drag the mouse across a region of interest while holding the ‘Alt’ key down (see Figure 11).

### c. Double click

A user can double click a region of interest to zoom in the tracks. A user can hold the ‘Shift’ button and double click the tracks to zoom out.

### d. Pan

When the tracks are zoomed in, a user can drag anywhere in the track panel to pan.



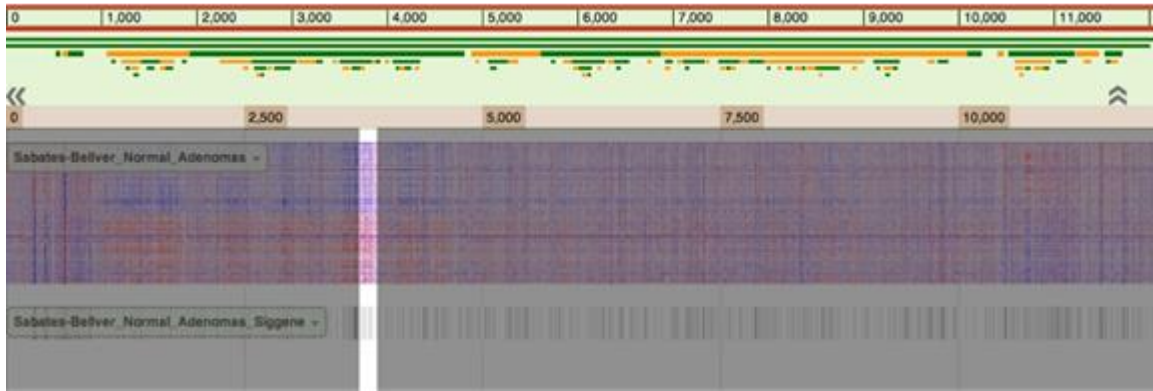


Figure 11. Zoom in the tracks by holding the ‘Alt’ key down and dragging the mouse.

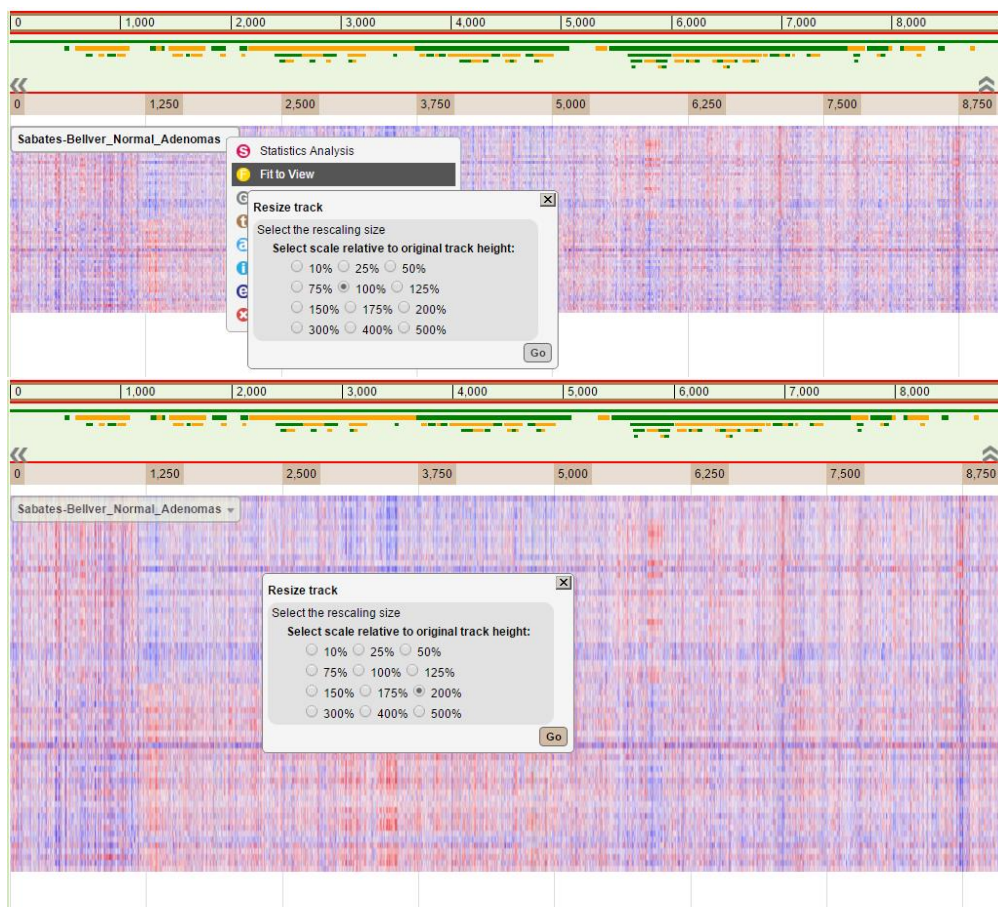


Figure 12. CCTs and CTBs can be vertically rescaled using the “Fit to View” option in the drop-down menu.

## 6. Resize tracks

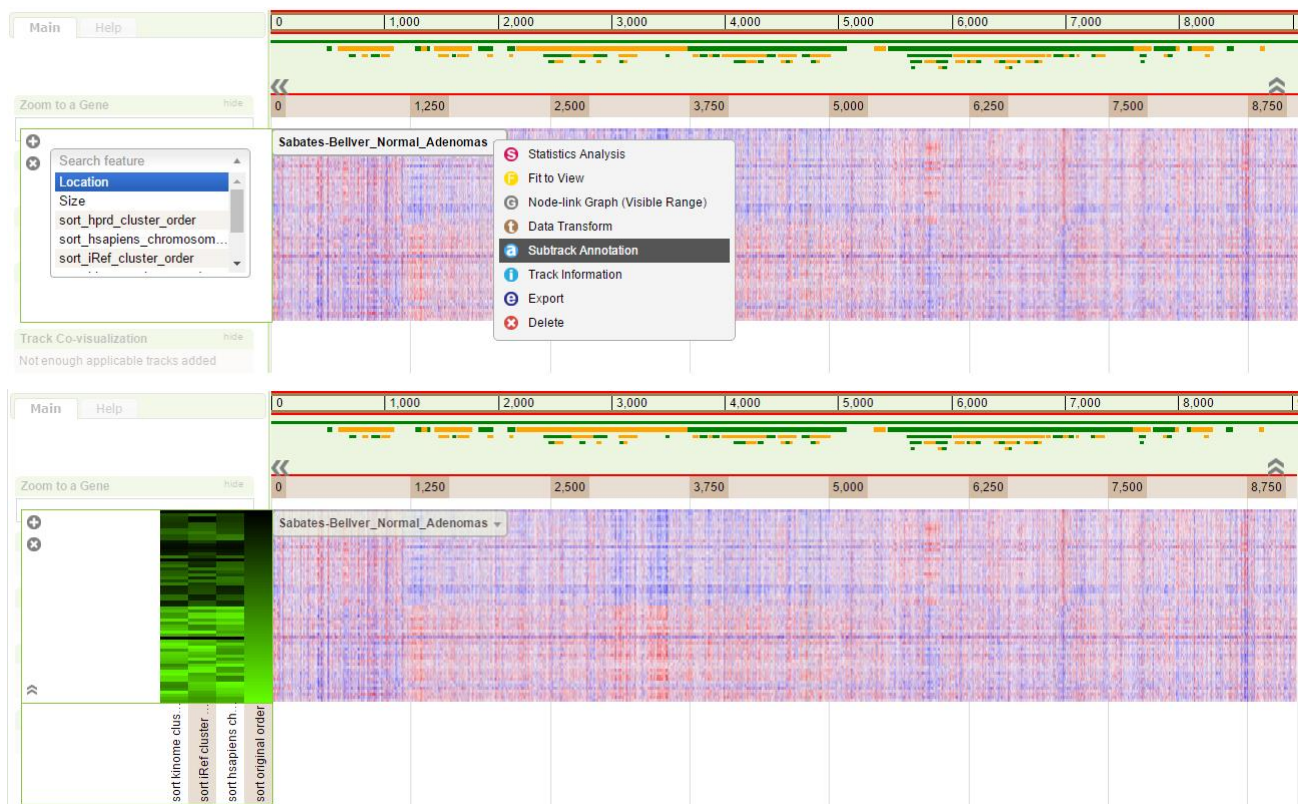
When CCT and CBT tracks contain a large number of samples, visualizing the entire track in your browser may not be possible without vertically scrolling the track space. The vertical rescaling option is available when you need to visualize an entire CCT or CBT on your screen. Select the “Fit to View” options under the track’s drop down menu, select a rescaling size, and

click Go. (See Figure 12).

## 7. Hierarchically cluster tracks

### a. Clustering w/respect to system views

For CCT and CBT tracks, the sample order based on hierarchical clustering is separately precalculated with respect to the genes found in each of preloaded system “views”. The clustering is done using the hclust library in R. The clustering sort order for each view is stored as a Subtrack Annotation feature (see Figure 13). When a track is initially loaded and the current view is a system (non-user uploaded) view, the default sorting for the track is the hierarchically clustered sort order for that particular view.

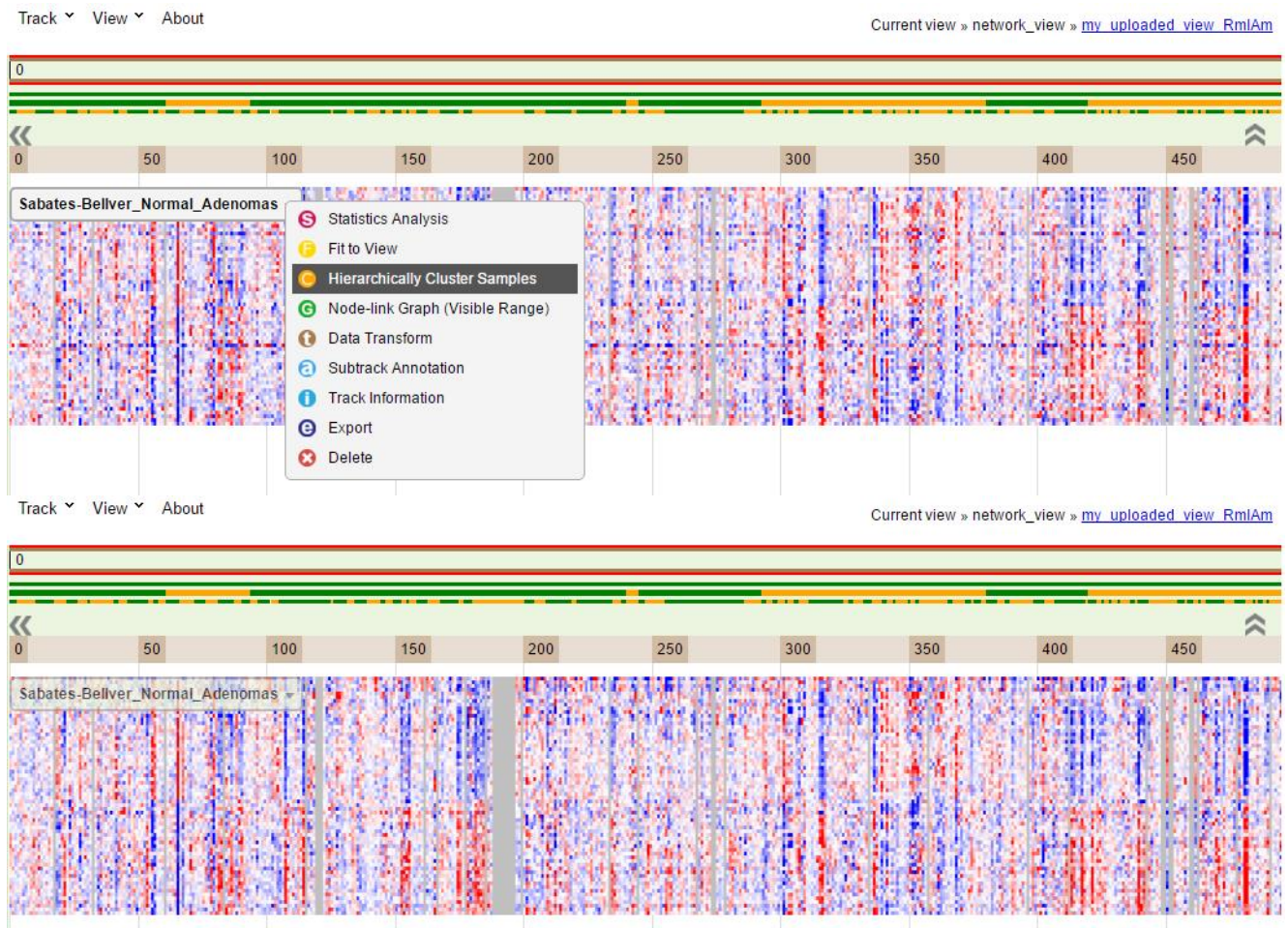


**Figure 13.** The sample orders for CCTs and CTBs can be resorted based on hierarchical clustering the samples with respect to the genes existing each of the preloaded system Views. Hierarchically clustered

### b. Clustering w/respect to user views

When the currently selected “view” is a user-uploaded view, the precalculated hierarchically clustered sample sort orders are not available. Instead, an on-the-fly clustering option will be available via the track’s drop-down menu (See Figure 14). This step might be slow, depending on the number of samples in your tracks and genes in the selected network.





**Figure 14.** When user networks are selected, the “Hierarchically Cluster Samples” option will be available in the drop down menus of CCTs and CTBs. This step might be slow, depending on the number of samples in your tracks and genes in your network.

## 8. Analyze tracks

The current version of NetGestalt provides eleven features to help users analyze the tracks.

### a. Network analysis

#### (i) Module enrichment

For SBTs or SCTs, NetGestalt can help users identify which modules (represented by bars with different length) are significantly correlated with the tracks.

For an SBT, NetGestalt uses the Fisher’s exact test to identify enriched modules identified from the active network. Enrichment p values are corrected for multiple comparisons by calculating the False Discovery Rates (FDRs).

Hovering the mouse over a track name, several buttons for track analysis will be shown in a drop-down menu (Figure 15). Clicking the “Network Analysis” button (red box in Figure 12), several options for network analysis will be shown. Choosing the “Module enrichment”

and clicking “Go” (blue box in Figure 15), the enriched network modules will be listed in a table in the “Enrichment Results” section located on the left panel of the page (red box in Figure 16). The user can click column title to sort the results by the corresponding column. The user can navigate through the pages by clicking the buttons right below the table or select the number of entries per page by clicking the drop-down menu at the bottom of the section (see purple box in Figure 16). Clicking one entry in the table will add overlapping genes between the enriched module and the original binary track as a new binary track (see green box in Figure 16) named by the user. The user can also add all enriched modules in a new composite track by clicking the “Add all related modules” link at the top of the table (see blue box in Figure 16). In this composite track, each row represents a hierarchical level of the network; enriched modules are colored in light red and genes in the original binary track are colored in red. The enrichment section can be hidden/shown by toggling the “hide”/ “show” button (see brown box in Figure 16).

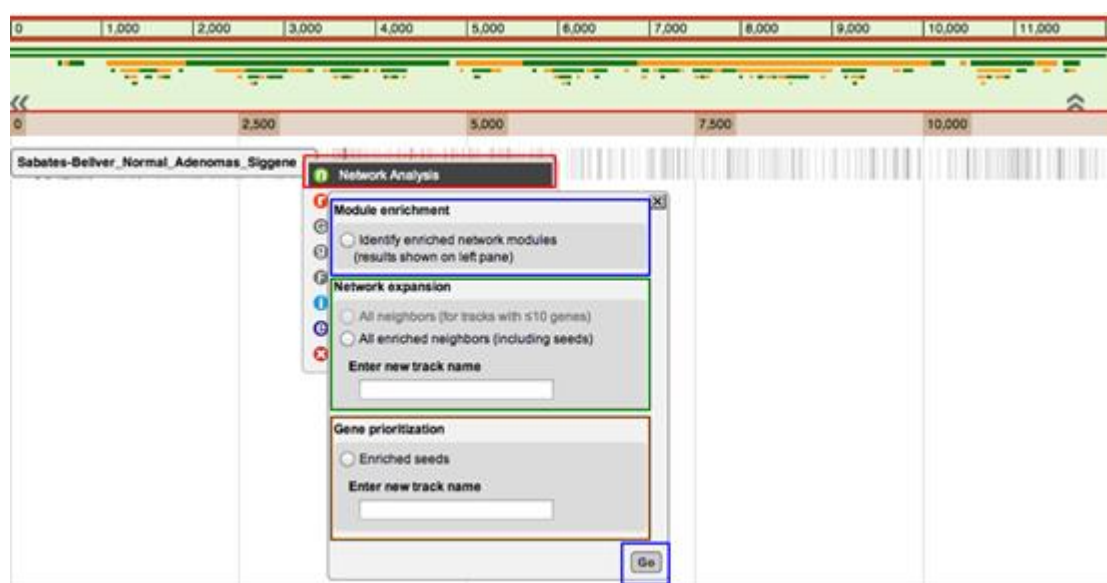


Figure 15. Network analysis in NetGestalt.

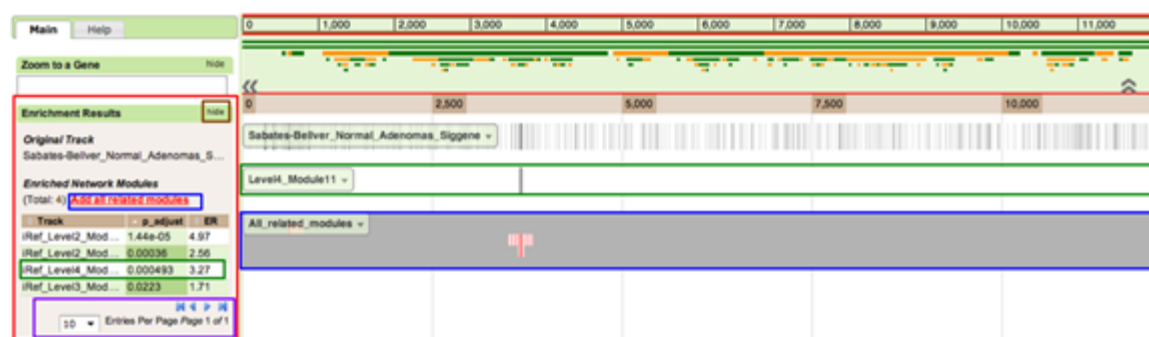


Figure 16. Output for network analysis in NetGestalt.

For an SCT, NetGestalt uses the Kolmogorov-Smirnov test (KS test) to identify enriched modules. When clicking an entry in the table (or “Add all related modules”),

NetGestalt will add the leading edge genes in the enriched module as a new binary track (see Subramanian et al. PNAS 102 (43): 15545-15550, 2005 for the definition of leading edge genes). Clicking the “Add all related modules” link at the top of the table will add a new composite track, in which each row represents a hierarchical level of the network; enriched modules are colored in light red and the leading edge genes are colored in red.

## **(ii) Network expansion**

To expand genes in an SBT (*i.e.*, the seed genes) to include other related genes in the network, two options are available in the Section “Network expansion” under “Network Analysis” (green box in Figure 15).

The “all neighbors” option works for SBTs containing 10 or fewer genes, and all direct neighbors of these genes in the network can be retrieved and shown in a new SBT together with the seed genes in the original SBT.

The “enriched neighbors” option works for any SBTs. Specifically, for each non-seed gene in the network, all direct neighbors of the gene are retrieved and evaluated for the enrichment of the seed genes using the Fisher’s exact test. All non-seed genes significantly enriched with seed neighbors according to a user defined FDR are identified and shown in a new SBT together with the seed genes in the original SBT.

## **(iii) Gene prioritization**

To prioritize genes in an SBT, NetGestalt provides a “Gene prioritization” feature under “Network Analysis” (brown box in Figure 15). Specifically, for each seed gene in the selected SBT, all direct neighbors of the gene are retrieved and evaluated for the enrichment of other seed genes using the Fisher’s exact test. Seed genes significantly enriched with other seed neighbors according to a user defined FDR are identified and shown in a new SBT.

## **b. Gene Set Enrichment**

NetGestalt compiles information from Gene Ontology (GO) and other five pathway databases including cell map, human cyc, kegg, nci-pid, and reactome to help users identify GO terms or pathways that are significantly correlated with an SCT or SBT. Figure 17 shows an enrichment result (red box) for a binary track based on GO biological process (BP) database. Enrichment analyses are based on the Fisher’s exact test and the KS test for SBTs and SCTs, respectively.

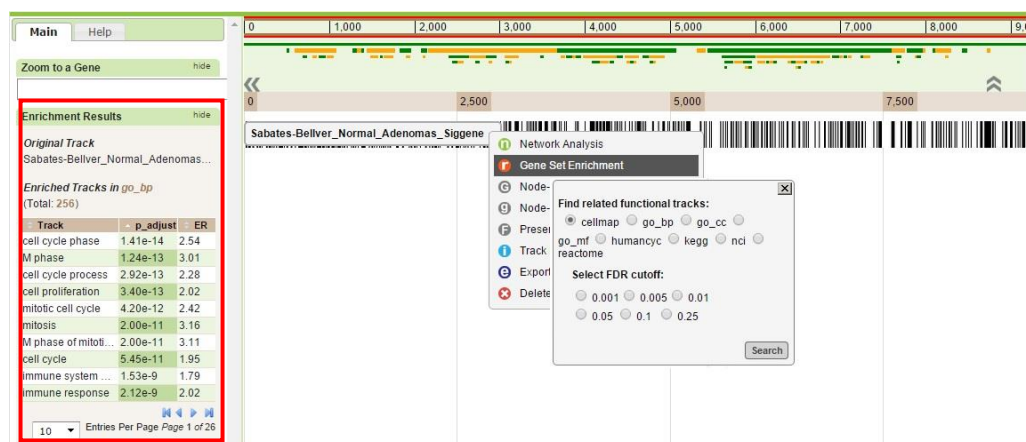


Figure 17. Gene set enrichment analysis in NetGestalt.

### c. Subtrack annotation

For composite tracks (CCTs and CBTs) containing multiple samples (*i.e.*, subtracks), using the “Subtrack Annotation” feature in the drop-down menu accessible from the track name, users can visualize sample information as a sample heat map with black to green colors for binary data (*e.g.*, Tissue), categorical data (*e.g.*, Location) and continuous data (*e.g.*, Size) (Figure 18). The order of the sample features can be rearranged, and the samples will be sorted according to the rightmost feature.

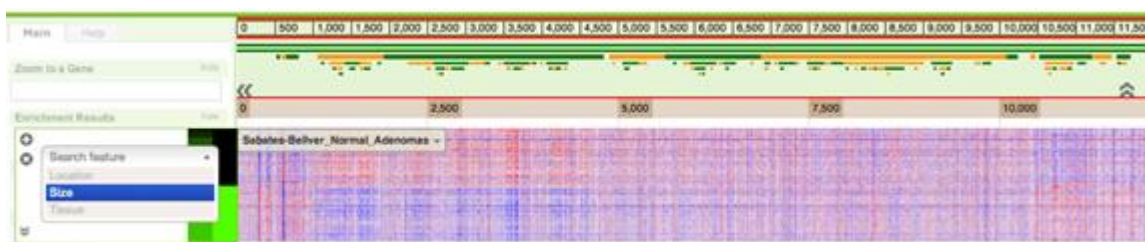


Figure 18. Subtrack annotation in NetGestalt.

### d. Statistical analysis

The users can conduct statistical association tests on CCTs or CBTs (*e.g.* clinical information like age, gender, etc) when subtrack annotation data is available (see section 6.c. above for more on subtrack annotation data). The “Statistical Analysis” button will be activated on a track’s drop-down menu if annotation data is available.

The statistical test used depends on the data type of the annotation data selected and the type of track data (whether it is a continuous or binary track).

### e. Data transformation

To better visualize the CCTs, NetGestalt provides a “Data Transform” feature, which allows users to perform gene-wise standardization by subtracting the gene-wise mean or median and set floor and ceiling values for the data (Figure 25). Similarly, using the “Data Transform” feature, users can also set floor and ceiling values for an SCT to improve data visualization.





Figure 25. Data transformation in NetGestalt.

#### f. Value-based filtering

The “Value-based Filtering” features allows users to filter for interesting genes (e.g. differentially expressed genes) from an SCT. After clicking the “Value-based Filtering” button in the menu (Figure 26), a filtering dialog will be displayed. The maximum and minimum values of the track are shown at the top of the dialog. The user can input the parameters to filter the track. After providing the name of the new track and clicking the “Add Track” button, an SBT will be added to the track viewing area (red box in Figure 26). Genes in the new SBT barplot are colored according to the original values in the SCT, with blue for negative values and red for positive values.

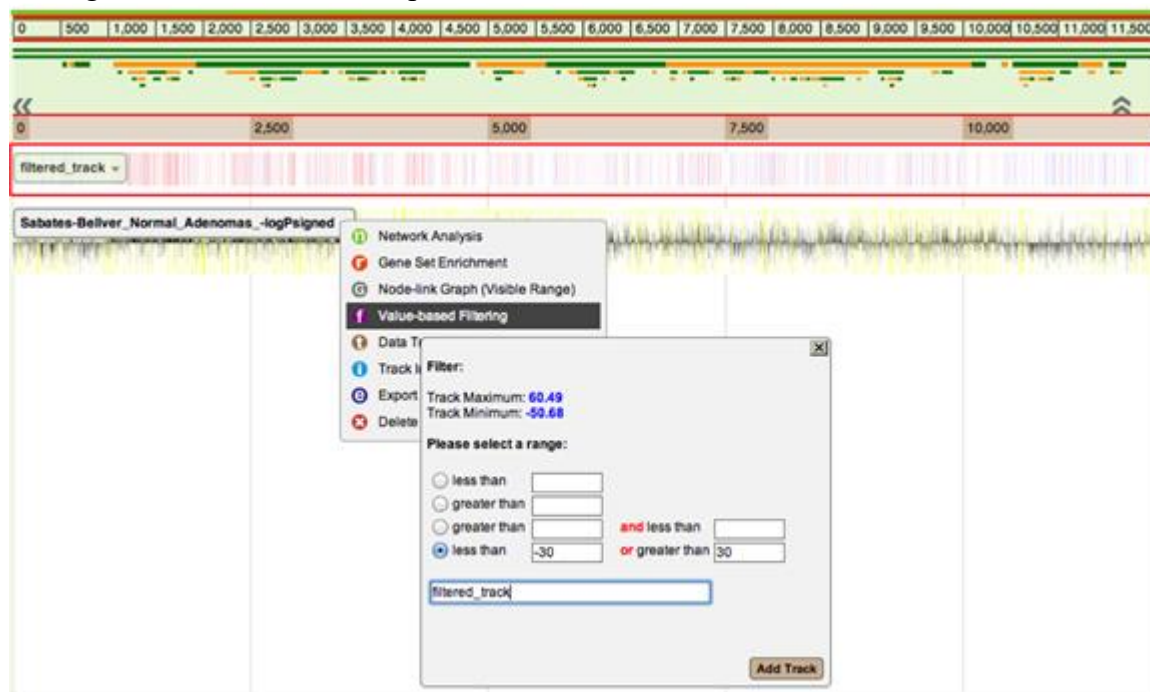


Figure 26. Value-based filtering in NetGestalt.

### g. Presence-based filtering

For SBTs, users can use the “Presence-based filtering” feature to focus on genes present in the current visible range. By clicking the “Presence-based Filtering” button (Figure 27), all genes present in the current visible range from the SBT will be identified, and data for these genes from all tracks in the track viewing area will be displayed in a new webpage (Figure 28). This feature is only activated when the number of genes in the current visible range from the SBT is between 2 to 100.

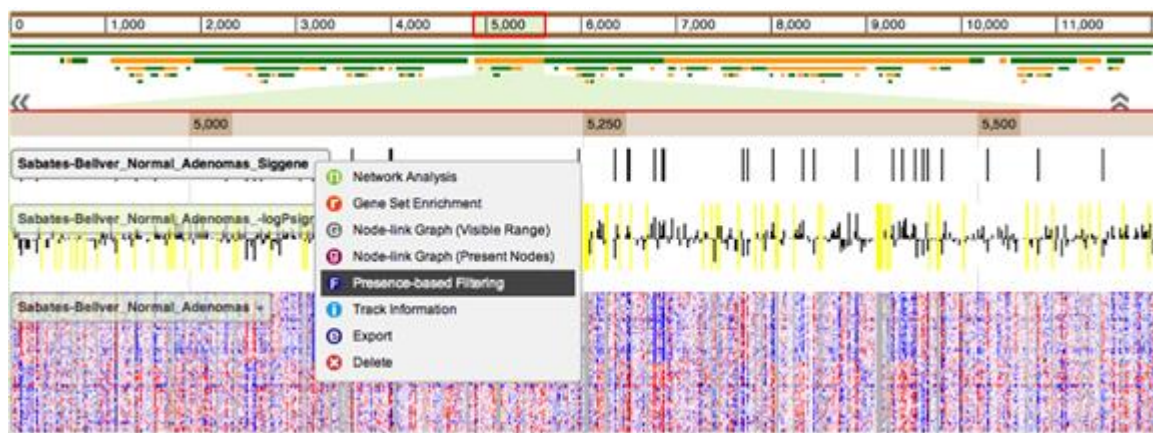


Figure 27. Presence-based filtering in NetGestalt.

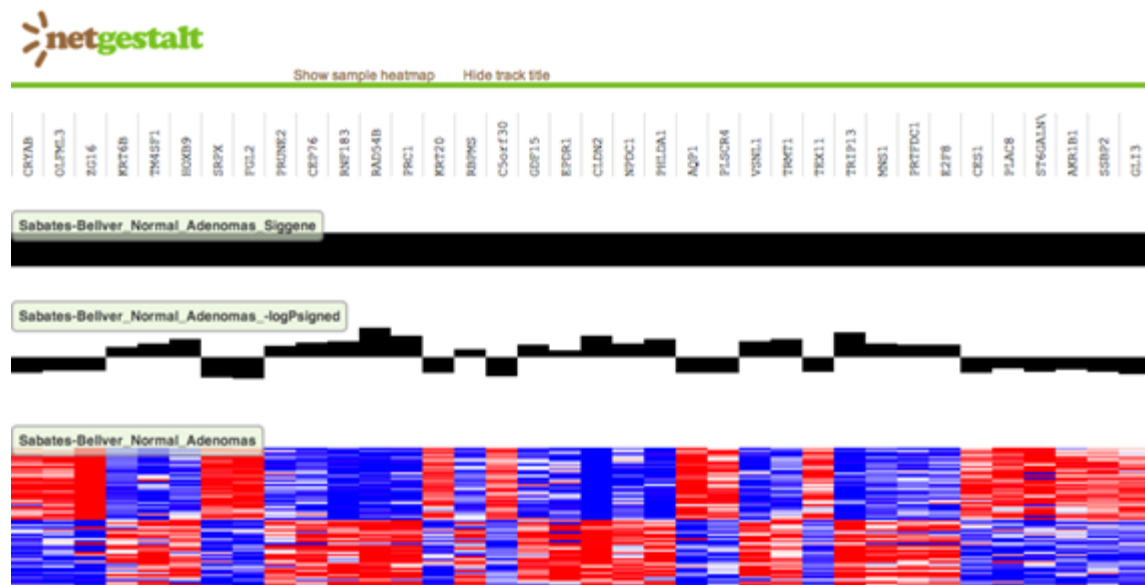


Figure 28. Output for presence-based filtering in NetGestalt.

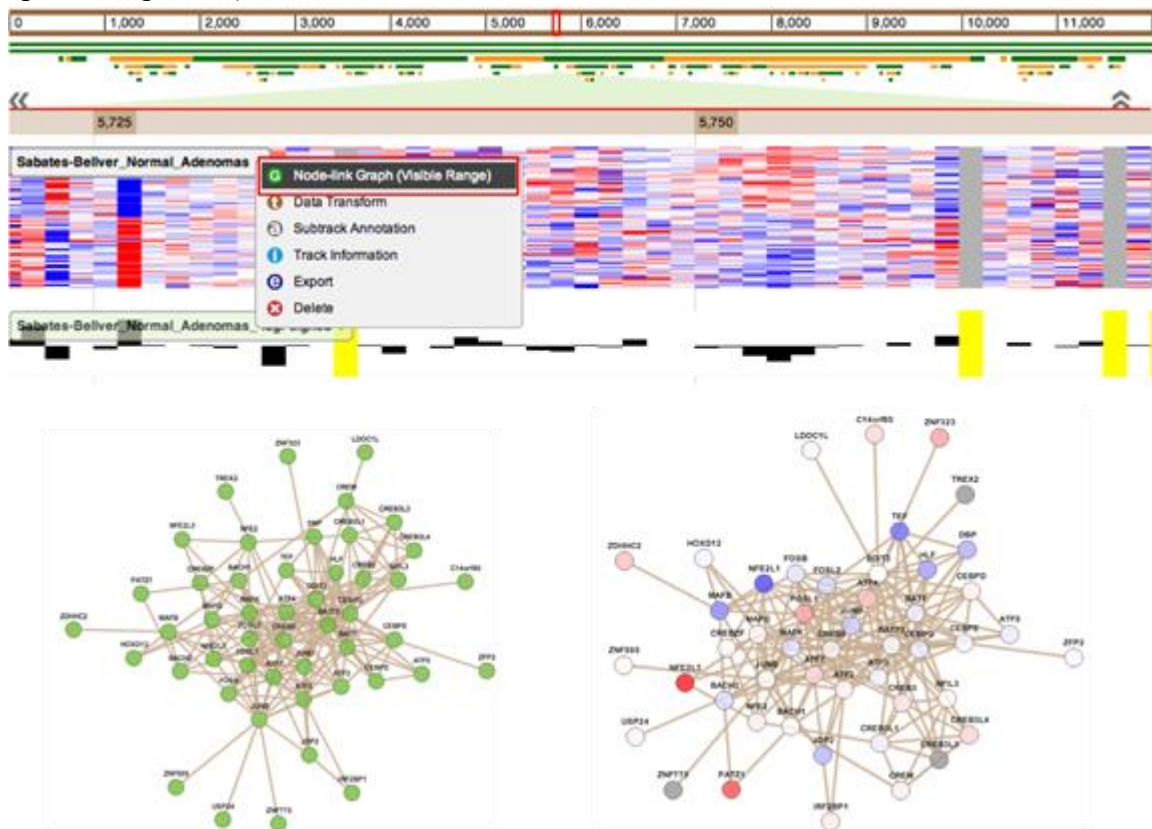
### h. Node-link Graph

NetGestalt also provides the traditional 2D network visualization using the Cytoscape Web plug in.

For CCTs and SCTs, when the number of genes in the current visible range is less than 500, the “Node-link Graph (Visible Range)” button in the drop-down menu will be activated (red box in the top plot of Figure 29). By clicking the “Node-link Graph (Visible Range)” button, network structure for all nodes in the current visible range will be displayed (left-bottom plot in Figure 29).

For an SCT, NetGestalt can color the genes in the network according to their original values in the SCT (right-bottom plot in Figure 29).

For SBTs, if genes in the track are colored, the corresponding genes in the network graph will also be colored with the same colors (left-bottom plot in Figure 30). In addition, by clicking the “Node-link Graph (Present Nodes)” button (red box in Figure 30), a network that only contains genes within the visible range and present in the SBT will be displayed (right-bottom plot in Figure 30).



**Figure 29. Visualize network structure for CCT and SCTs in NetGestalt.**

#### **i. Zoom to a gene**

The users can use the “Zoom to a gene” feature to zoom directly to a gene. When typing the gene symbol in the “Zoom to a gene” box (see purple box in Figure 31), NetGestalt will list all matching symbols below the box on the fly. After selecting a symbol, NetGestalt will zoom in to a small region containing the selected symbol and use a vertical red line to highlight the gene in all tracks in the track viewing area. (Figure 31).

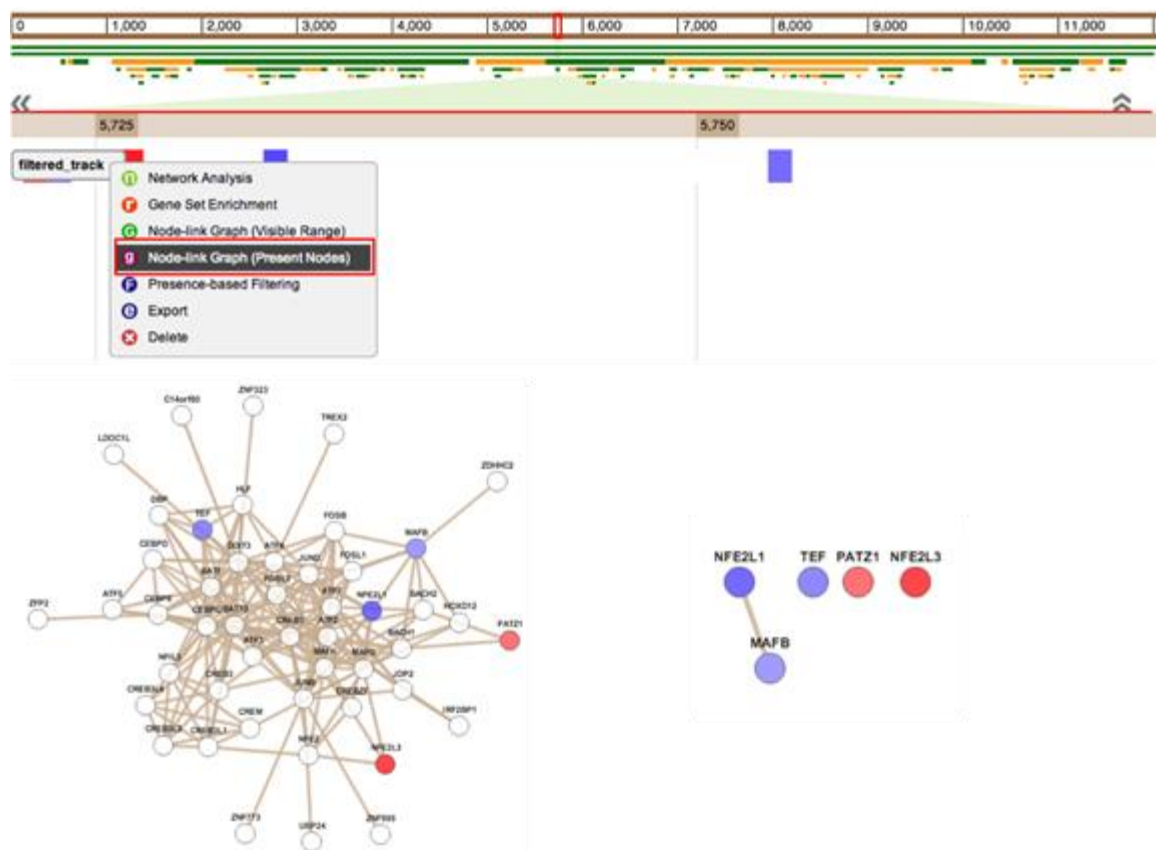


Figure 30. Visualize network structure for SBT in NetGestalt.

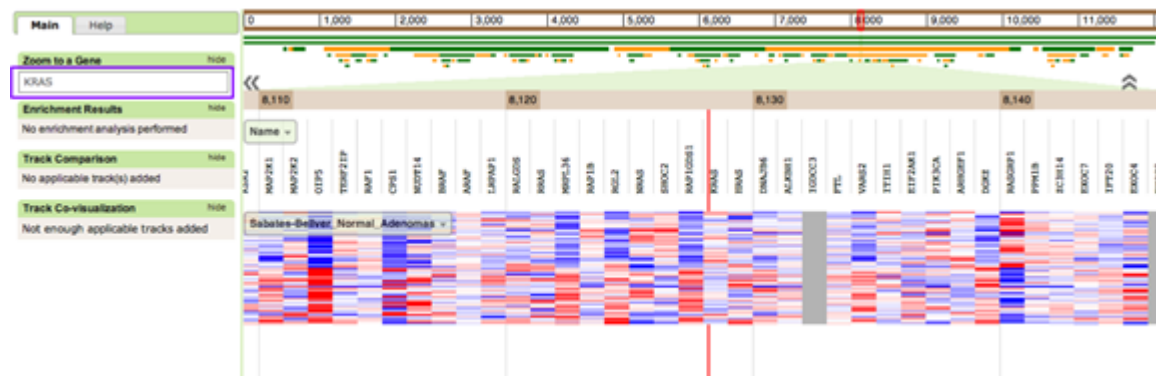


Figure 31. Zoom to a gene in NetGestalt.

## j. Track comparison

NetGestalt uses an interactive Venn diagram to help users compare different SBTs. The user can first select the binary tracks from the “Track Comparison” section located in the left panel of the page (see brown box in Figure 32). An SBT automatically appears in this section when it is added to the track viewing area and will be removed from the section after it is deleted. At most three tracks can be compared at the same time. As soon as tracks are selected, a clickable Venn diagram will be shown below the track names (Figure 32). To help users



distinguish different binary tracks easily, NetGestalt uses the same color for the selected track name in the “Track Comparison” section, circle in the Venn diagram and upper and lower borders of the binary track visualized in the viewing area. Clicking each part of the Venn diagram will add a new binary track to the track viewing area. For example, the user can click the overlapping part of the Venn diagram and add a new binary track. User can also click the blank region of the Venn diagram to add the union of the genes. Genes in the new binary track are colored corresponding to the colors in Venn diagram.

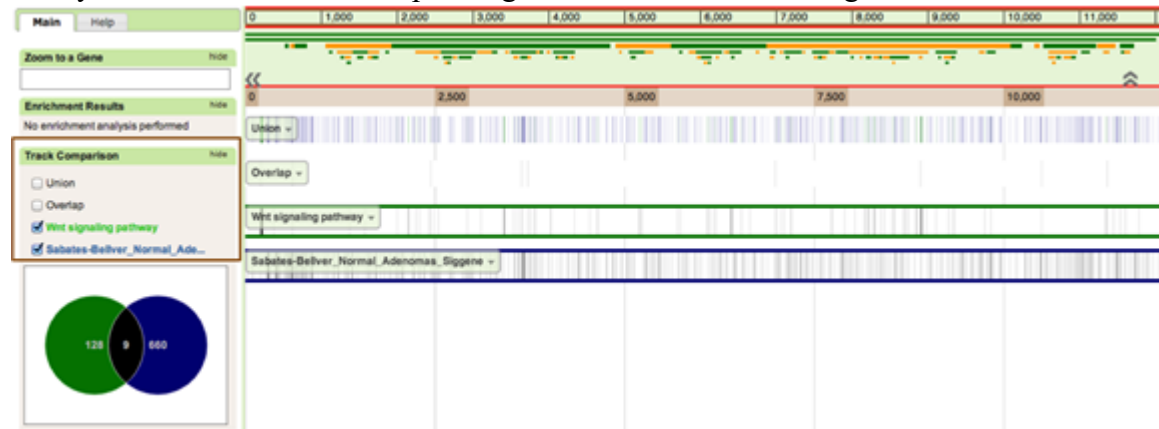



Figure 32. Track comparison in NetGestalt.

#### k. Track co-visualization

NetGestalt allows users to co-visualize two single tracks (SBT or SCT) in a node-link graph, using the border and fill colors of the nodes to represent data in the two tracks, respectively. First, the user should select the tracks to be co-visualized (SBT or SCT) and the node attributes (border or fill colors) associated with each track in the “Track Co-visualization” section located in the left panel of the page (brown box in Figure 33). After clicking the “G” or “g” buttons, a node-link graph, which contains edges between all genes in the visible range or all present genes in the visible range, will be displayed (left-bottom and right-bottom plots in Figure 33).

#### l. Switch between different views

The users can visualize the same set of tracks in different network views by making changes using the “Views”  “Select” menu as shown in the Figure 2. This feature allows users to explore the same data sets in the different biological contexts.

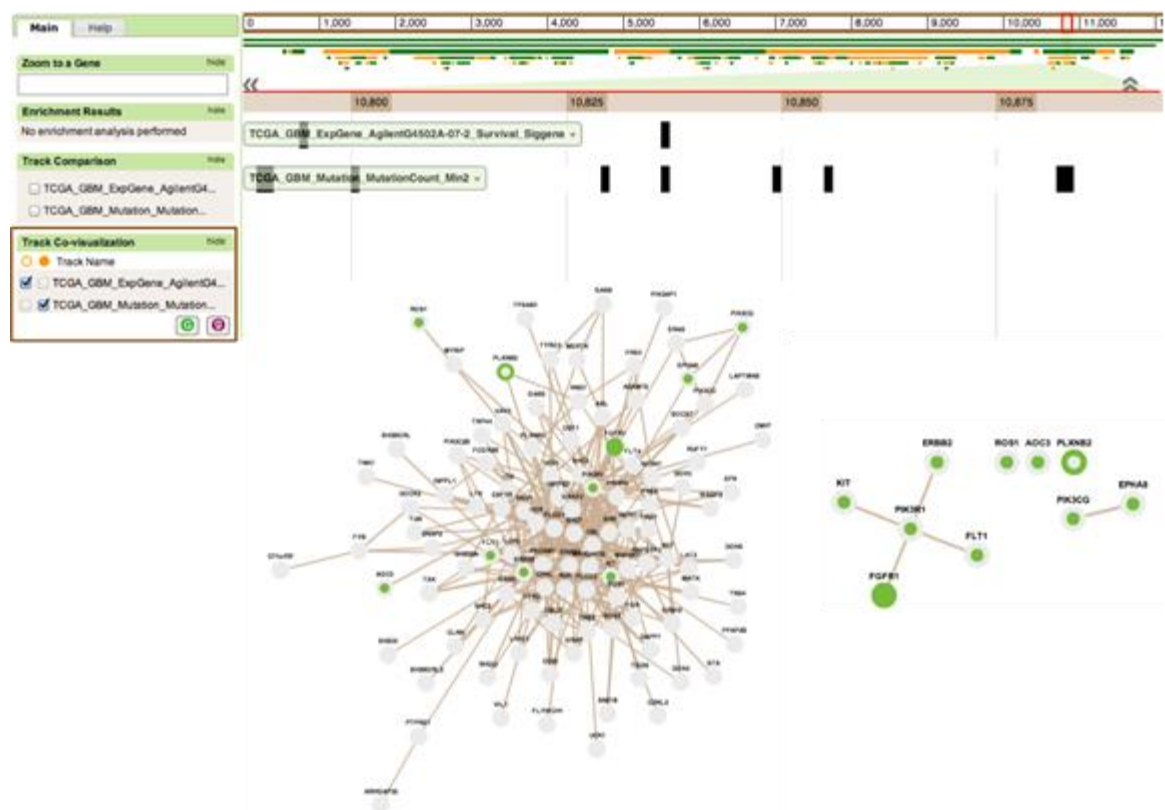


Figure 33. Track co-visualization in NetGestalt.

### III. Portal descriptions

There are multiple portals available for NetGestalt. Each portal contains both the protein-protein interaction network “views” for a given species along with a chromosome “view”. Currently, there are separate portals for the following species: *Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Danio rerio*. Each portal comes preloaded with tracks. For the default human portal, the context is generic, with preloaded track limited to some example tracks, as well as functional tracks from sources like KEGG, Reactome, GO, and DrugBank. The rest of the species-portals come preloaded with just GO functional tracks.

Two portals, the Human: Colorectal Cancer (CRC) portal and the Human: Clinical Proteomics Tumor Analysis Consortium (CPTAC) portal both come preloaded with tracks containing the experimental results from specific studies. The data sources and data processing methods for these portals are described below.

#### 1. Generating Views

Each portal contains at least one view derived from a protein-protein interaction (PPI) network and one viewed based on chromosomal locations. See section II.2.a for more on chromosome view vs network views.

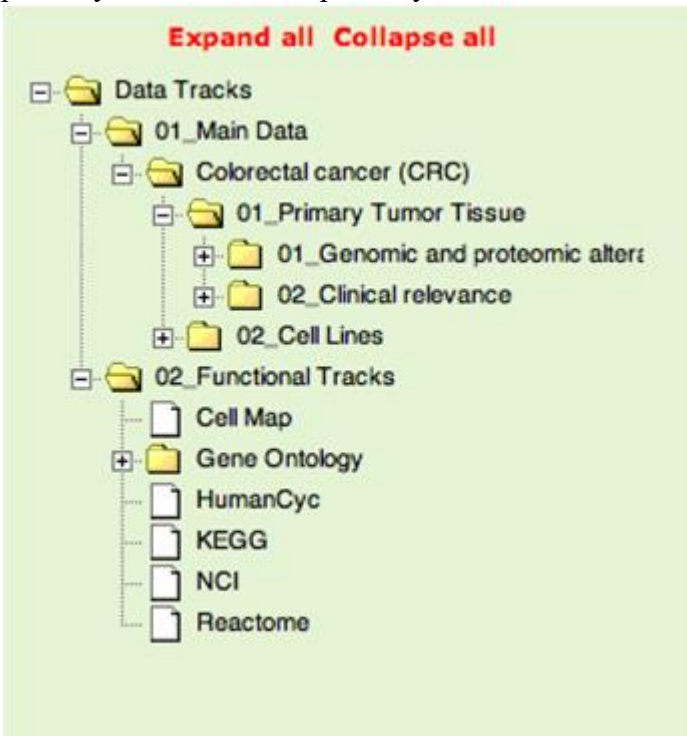
Currently, the human portals (CRC, CPTAC, and default human portals) contain five network views: “hprd”, corresponding to the HPRD human protein-protein interaction (PPI) network (<http://www.hprd.org/>), “iRef”, corresponding to the iRef human PPI network (<http://wodaklab.org/iRefWeb/>), “kinome”, corresponding to the human kinome data from the Kinome Renderer (Chartier, M., et al, 2013), “TFClass\_human”, corresponding to the human transcription factors from TFClass (Wingender, E., et al, 2013), and “phosphatase\_human”, corresponding to the human phosphatases (Liberti, S, 2013). For the non-human portals, the protein-protein interaction data of different organisms from BioGrid was processed with the NetSAM package to create a PPI view. Additionally, the Mouse portal also contains a “TFClass\_mouse” view corresponding to the mouse transcription factors from TFClass (Wingender, E., et al, 2013). See section II.2.b for more on NetSAM.

For the chromosome views, the chromosome information for eight organisms was downloaded from BioMart. We generated the chromosome view based on the gene positions in each chromosome and the band information. The first level in the view contained all genes in all chromosomes and the second level was the genes in each chromosome. If the organism contained the band information, the bands of each chromosome will be shown from the third level. The gene positions in the view were determined by the gene positions in the chromosome.

#### 2. The Colorectal Cancer (CRC) portal track description

As shown in Figure 34, the current version of NetGestalt CRC portal contains: (1) genomic, epigenomic, transcriptomic, proteomic, and clinical data for The Cancer

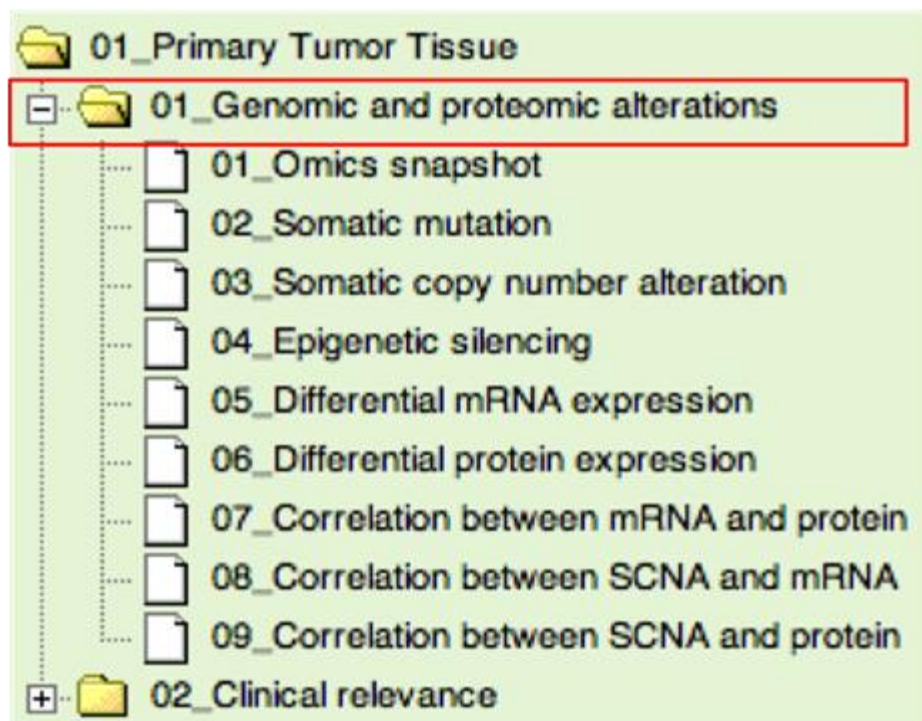
Genome Atlas (TCGA) CRC cohort; (2) mRNA expression data from several Gene Expression Omnibus (GEO) CRC cohorts that come with survival information; (3) data from CRC cell lines including drug response, genomic, and transcriptomic data from the Cancer Cell Line Encyclopedia (CCLE) project and shRNA screen data from the Achilles project; and (4) functional tracks including Cell Map pathways, GO Biological Processes, GO Cellular Components and GO Molecular Functions, HumanCyc pathways, KEGG pathways, NCI pathways and Reactome pathways.



**Figure 34. Data tracks in NetGestalt CRC portal.**

#### **a. Genomic and proteomic alterations in the TCGA CRC tumor cohort**

Tracks derived from the multidimensional omics data on the TCGA CRC tumor cohort are included in the Category "01\_Genomic and proteomic alterations" (red box in Figure 35). These tracks are divided into nine sub-categories: 01\_Omics snapshot, 02\_Somatic mutation, 03\_Somatic copy number alteration (SCNA), 04\_Epigenetic silencing, 05\_Differential mRNA expression, 06\_Differential protein expression, 07\_Correlation between mRNA and protein, 08\_Correlation between SCNA and mRNA and 09\_Correlation between SCNA and protein.



**Figure 35. Genomic and proteomic alteration based on CRC tumor tissue.**

Sub-category 01\_Omics snapshot only contains one track "Omics snapshot", which shows the summary of alterations (somatic mutation, somatic copy number alterations, epigenetic alterations, differential expression at mRNA level and protein level) for all the genes based on data from the TCGA CRC tumor cohort.

Sub-category 02\_Somatic mutation contains four tracks, including one CBT file recording the binary mutation matrix, two SBT files recording the significantly mutated genes and genes mutated in at least 5% of all the CRC samples and one SCT file recording mutation counts in log scale, i.e.  $\log_2(\text{mutation count}+1)$ .

Sub-category 03\_Somatic copy number alteration (SCNA) contains four tracks, including two CCT files recording the gene level SCNA matrix and focal SCNA matrix, two SBT files recording genes in the focal amplification regions and focal deletion regions.

Sub-category 04\_Epigenetic silencing contains two tracks, including one CCT file recording the methylation matrix and one SBT file recording the candidate epigenetically silenced genes.

Sub-category 05\_Differential mRNA expression contains six tracks, including one CCT file recording the gene expression matrix, three SCT files recording the t-statistic values, signed  $-\log P$  (p values were calculated based on t-test) and  $\log_2$  fold changes and two SBT files recording the up-regulated and down-regulated differentially expressed genes.

Sub-category 06\_Differential protein expression contains six tracks, including one CCT file

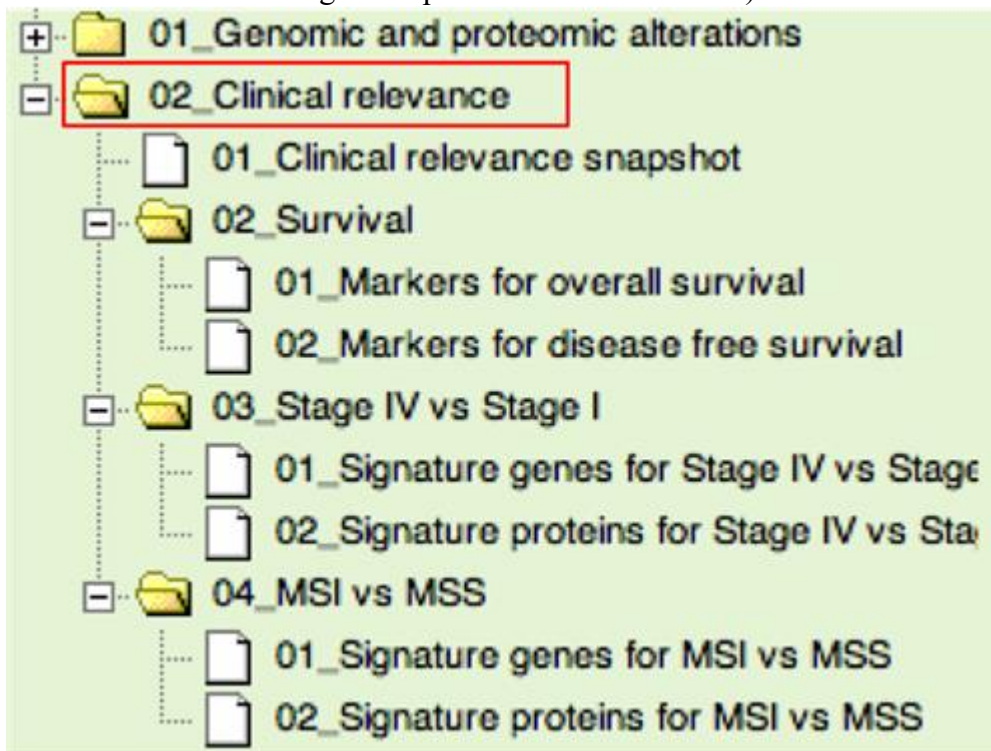


recording the protein expression matrix, three SCT files recording the W-statistic values, signed -logP (p values were calculated based on Wilcoxon test) and log2 fold changes and two SBT files recording the up-regulated and down-regulated differentially expressed proteins.

For 07\_Correlation between mRNA and protein, 08\_Correlation between SCNA and mRNA and 09\_Correlation between SCNA and protein, each of them contains two tracks, both of which are SCT files. These SCT files record the Spearman's correlation coefficient and corresponding signed -logp.

#### **b. Clinical relevance based on CRC tumor tissue**

All the clinical relevant data tracks are included in the Category "02\_Clinical relevance" (red box in Figure 36). These tracks are divided into four categories: 01\_Clinical relevance snapshot, 02\_Survival (including Markers for overall survival and Markers for disease free survival), 03\_Stage IV vs Stage I (including Signature genes for Stage IV vs Stage I and Signature proteins for Stage IV vs Stage I), 04\_MSI vs MSS (including Signature genes for MSI vs MSS and Signature proteins for MSI vs MSS).



**Figure 36. Clinical relevance based on CRC tumor tissue.**

Sub-category 01\_Clinical relevance snapshot only contains one track "Clinical\_relevance\_snapshot", which shows the summary of clinical relevance including markers for disease free survival, markers for overall survival, signature genes for MSI vs MSS and signature genes for stage IV vs stage I for all the genes based on multiple datasets.

Sub-category 02\_Survival contains two lower level sub-categories 01\_Markers for overall survival and 02\_Markers for disease free survival. Each of them contains eleven tracks, including (1) eight SCT files recording the correlation (signed  $-\log p$  and  $\log_2(\text{Hazard Ratio})$ ) between gene expression and survival (overall survival or disease free survival) based on Cox regression model and four datasets (two tracks for each of the four datasets); (2) three summary tracks (one continuous track and two binary tracks) summarizing the results based on the four datasets by order statistics (Wang, et al., 2013).

Sub-category 03\_Stage IV vs Stage I contains two lower level sub-categories 01\_Signature genes for Stage IV vs Stage I and 02\_Signature proteins for Stage IV vs Stage I. 01\_Signature genes for Stage IV vs Stage I contains fifteen tracks, including (1) twelve SCT files recording the differential expression (t-statistic, signed  $-\log p$  and  $\log_2(\text{Fold Change})$ ) of genes comparing Stage IV with Stage I CRC samples based on t test and four datasets (three tracks for each of the four datasets); (2) three summary tracks (one continuous track and two binary tracks) summarizing the results based on the four datasets by order statistics. 02\_Signature proteins for Stage IV vs Stage I contains only three tracks recording the differential expression (W-Statistic, signed  $-\log p$  and  $\log_2(\text{Fold Change})$ ) of proteins comparing Stage IV with Stage I CRC samples based on Wilcoxon test and protein data from (Zhang, et al., 2014).

Sub-category 04\_MSI vs MSS contains two lower level sub-categories 01\_Signature genes for MSI vs MSS and 02\_Signature proteins for MSI vs MSS. 01\_Signature genes for MSI vs MSS contains fifteen tracks, including (1) twelve SCT files recording the differential expression (t-statistic, signed  $-\log p$  and  $\log_2(\text{Fold Change})$ ) of genes comparing MSI with MSS CRC samples based on t test and four datasets (three tracks for each of the four datasets); (2) three summary tracks (one continuous track and two binary tracks) summarizing the results based on the four datasets by order statistics. 02\_Signature proteins for MSI vs MSS contains only three tracks recording the differential expression (W-statistic, signed  $-\log p$  and  $\log_2(\text{Fold Change})$ ) of proteins comparing MSI with MSS CRC samples based on Wilcoxon test and protein data from (Zhang, et al., 2014).

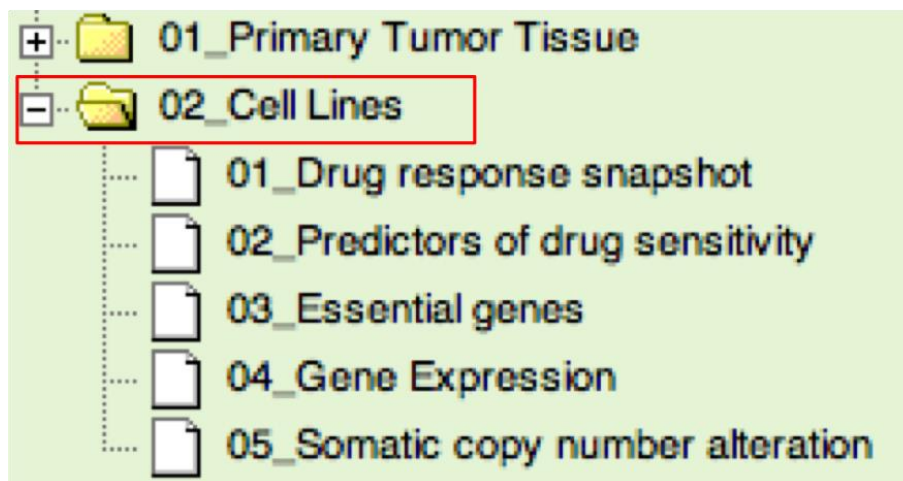


Figure 37. Clinical relevance based on CRC tumor tissue.

### c. Tracks based on CRC cell lines

All the data tracks derived from CRC cell lines are included in the Category "02\_Cell Lines" (red box in Figure 37). These tracks are divided into five categories: 01\_Drug response snapshot, 02\_Predictors of drug sensitivity, 03\_Essential genes, 04\_Gene Expression and 05\_Somatic copy number alteration.

Category 01\_Drug response snapshot only contains one track "Drug\_response\_snapshot", which shows the Spearman's correlation coefficient between response (activity area) of the 24 compounds and mRNA expression of all the genes.

Category 02\_Predictors of drug sensitivity contains 24 tracks recording the Spearman's correlation coefficient between response (activity area) of 24 compounds and mRNA expression of all the genes separately.

Category 03\_Essential genes contains one track "Achilles\_RNAi\_CRC\_Cellline\_specific\_Essential\_Genes", which shows the CRC cell line specific essential genes from Project Achilles (Cheung, et al., 2011).

Category 04\_Gene Expression contains one track "CCLE\_CRC\_ExpGene\_AffyU133\_2", which shows the gene expression matrix of CRC cell lines from Cancer Cell Line Encyclopedia study.

Category 05\_Somatic copy number alteration contains one track "CCLE\_CRC\_CNA\_AffySNP6", which shows the copy number alteration matrix of CRC cell lines from Cancer Cell Line Encyclopedia study.

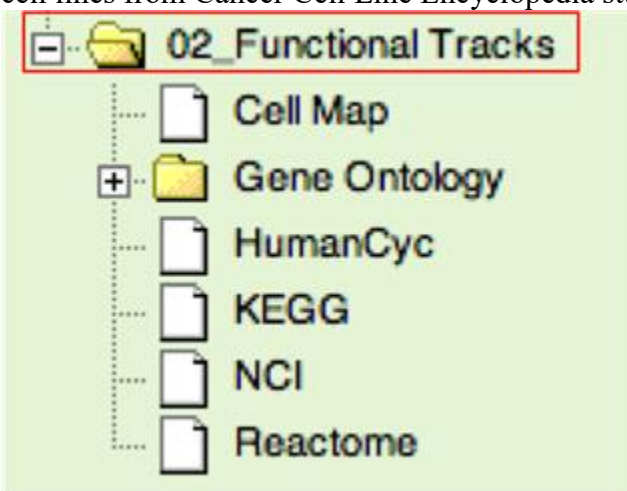


Figure 38. Functional tracks.

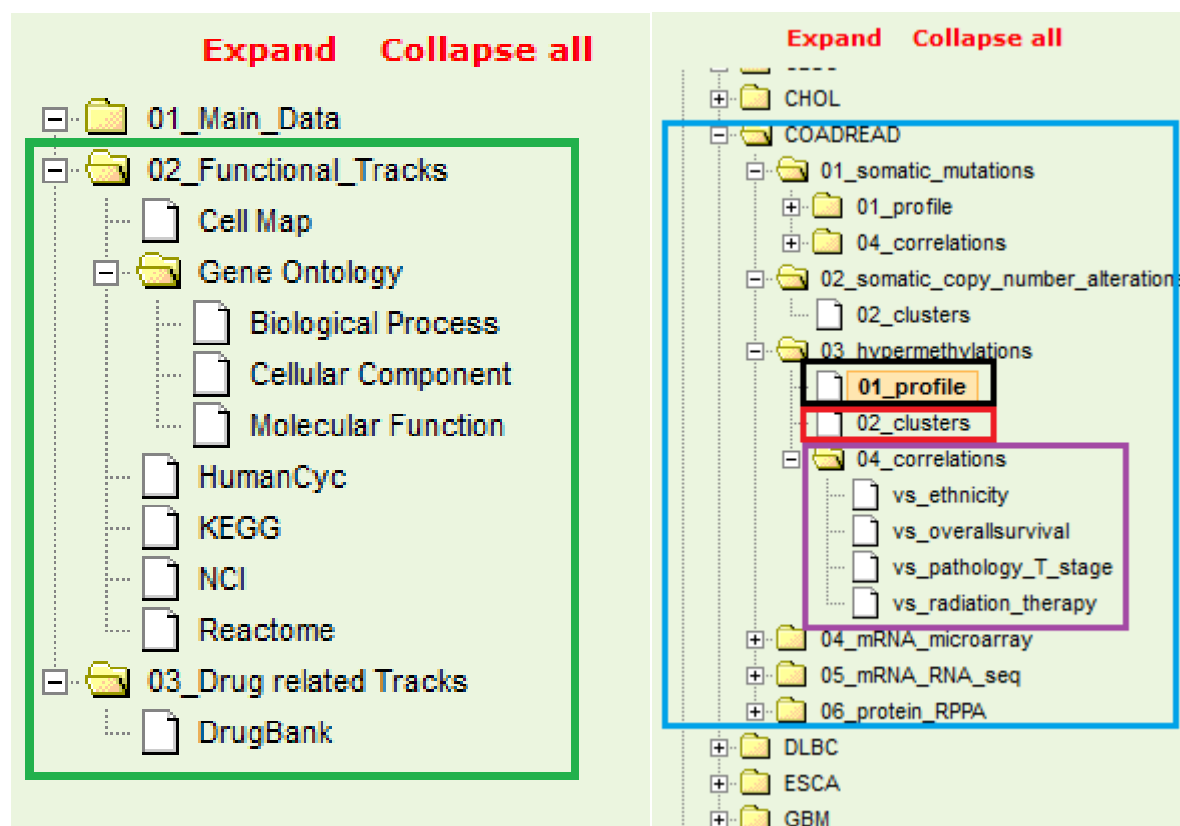
### d. Functional tracks

All the functional tracks are included in the Category "02\_Functional Tracks" (red box in Figure 38). These tracks are from six databases: Cancer Cell Map (10 pathways), Gene Ontology (808 Biological Processes, 231 Cellular Components and 383 Molecular Functions), HumanCyc (267 pathways), KEGG (200 pathways), NCI (223 pathways) and Reactome (1108 pathways).



### 3. The Cancer Genome Atlas (TCGA) Portal

As shown in Figure 39, the current version of NetGestalt TCGA portal contains: (1) proteomic, genomic, epigenomic, transcriptomic, and clinical data from The Cancer Genome Atlas (TCGA) study (blue box in Figure A); (2) Clustering analysis results from the Broad Institute (red box in Figure A); (3) statistical correlation analysis results (purple box in Figure A); and (4) functional tracks including Cell Map pathways, GO Biological Processes, GO Cellular Components and GO Molecular Functions, HumanCyc pathways, KEGG pathways, NCI pathways and Reactome pathways (green box in Figure 39).



**Figure 39.** Overview of track types available on the TCGA portal. Functional tracks containing SBTs of genes associated with Gene Ontology terms or other biological databases and DrugBank target genes are highlighted in the green box. The colorectal (COADREAD) tracks, found in the “01\_Main\_data” branch, are shown in the blue box. If available, tracks were generated from the TCGA data for somatic mutations, somatic copy number alternations, hypermethylations, mRNA microarray, RNAseq, and RPPA data. Data sets are further divided between profile tracks (clinically annotated matrices of the experimental data, shown in black box), clustering tracks (separate SBTs for each cluster for the various clustering analysis results, shown in the red box), and clinical correlation tracks which are further separated by the clinical data tested against in the correlation analysis. (purple box).

Data on the follow tissues is available: Acute Myeloid Leukemia (LAML), Adrenocortical Carcinoma (ACC), Bladder Urothelial Carcinoma (BLCA), Brain Lower Grade Glioma(LGG), Breast Invasive Carcinoma (BRCA), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), Cholangiocarcinoma (CHOL), Colon Adenocarcinoma (COAD),

Colorectal Adenocarcinoma (COADREAD), Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC), Esophageal Carcinoma (ESCA), Glioma (GBMLGG), Glioblastoma Multiforme (GBM), Head and Neck Squamous Cell Carcinoma (HNSC), Kidney Chromophobe (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney Renal Papillary Cell Carcinoma (KIRP), Pan-kidney cohort (KICH + KIRC + KIRP) (KIPAN), Liver Hepatocellular Carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Mesothelioma (MESO), Ovarian Serous Cystadenocarcinoma (OV), Pancreatic Adenocarcinoma (PAAD), Pheochromocytoma and Paraganglioma (PCPG), Prostate Adenocarcinoma (PRAD), Rectum Adenocarcinoma (READ), Sarcoma (SARC), Skin Cutaneous Melanoma (SKCM), Stomach Adenocarcinoma (STAD), Stomach and Esophageal carcinoma (STES), Testicular Germ Cell Tumors (TGCT), Thymoma (THYM), Thyroid Adenocarcinoma (THCA), Uterine Corpus Endometrioid Carcinoma (UCEC), Uterine Carcinosarcoma (UCS), Uveal Melanoma (UVM).

#### **a. TCGA profile tracks**

The TCGA profile tracks are generated from the TCGA data for somatic mutations, somatic copy number alternations, hypermethyations, mRNA microarray, RNAseq, and RPPA if the data is available. In the current version of the CTPAC portal, the TCGA data is taken from the 01/28/2016 version of the TCGA result sets.

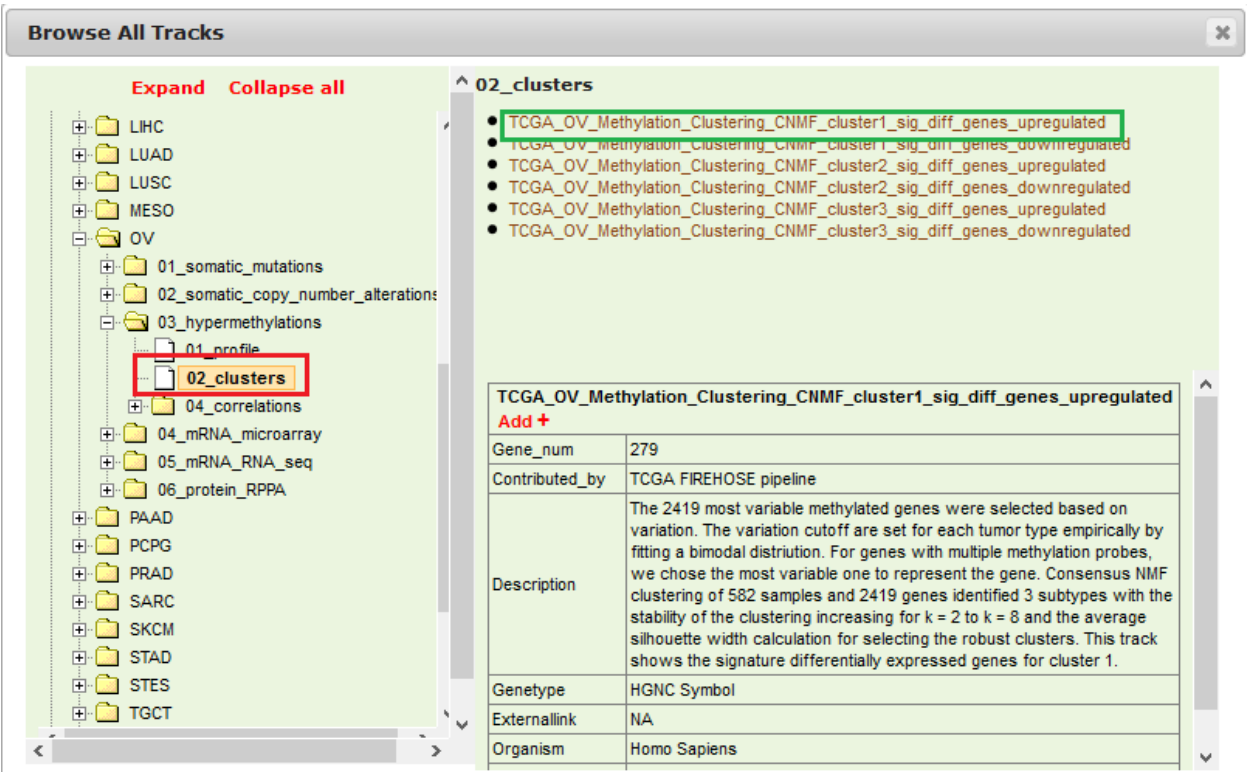
The following data sets from TCGA were added to the CTPAC portal:

1. Somatic mutation data:
  - i. Mutation data was obtained from the MutSig analysis results data sets from TCGA (e.g. [http://gdac.broadinstitute.org/runs/analyses\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD-TP.MutSigNozzleReport2CV.Level\\_4.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/analyses_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD-TP.MutSigNozzleReport2CV.Level_4.2015040200.0.0.tar.gz))
  - ii. Silent or non-somatic mutations are filtered out.
  - iii. CBTs are generated for each version of MutSig available (version 1.5, 2.0, CV, and 2CV), where 1 = mutation present, 0 = mutation not present.
2. mRNA transcriptome microarray data:
  - i. Lowess-normalized log2-transformed transcriptome data was obtained from TCGA (e.g. [http://gdac.broadinstitute.org/runs/stddata\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD.Merge\\_transcriptome\\_agilentg4502a\\_07\\_3\\_unc\\_edu\\_Level\\_3\\_unc\\_lowess\\_normalization\\_gene\\_level\\_data.Level\\_3.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/stddata_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD.Merge_transcriptome_agilentg4502a_07_3_unc_edu_Level_3_unc_lowess_normalization_gene_level_data.Level_3.2015040200.0.0.tar.gz) )
3. RNAseq data:
  - i. Normalized RSEM gene-level RNAseq data was obtained from the HiSeq rnaseq (v2) normalized gene-level data sets (e.g. [http://gdac.broadinstitute.org/runs/stddata\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD.Merge\\_rnaseqv2\\_illuminahi\\_seq\\_rnaseqv2\\_unc\\_edu\\_Level\\_3\\_RSEM\\_genes\\_normalized\\_data.Level\\_3.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/stddata_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD.Merge_rnaseqv2_illuminahi_seq_rnaseqv2_unc_edu_Level_3_RSEM_genes_normalized_data.Level_3.2015040200.0.0.tar.gz) )
  - ii. The “normalized count” values provided by TCGA are  $\log_2(\text{value} + 1)$  transformed and generated as a CCT.

4. Hypermethylation data:
  - i. Methylation data was obtained from the Human Methylation 450 TCGA data set (e.g.  
[http://gdac.broadinstitute.org/runs/stddata\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD.Merge\\_methylation\\_human\\_methylation450\\_jhu usc edu\\_Level\\_3\\_within\\_bioassay\\_data\\_set\\_function\\_data.Level\\_3.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/stddata_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD.Merge_methylation_human_methylation450_jhu usc edu_Level_3_within_bioassay_data_set_function_data.Level_3.2015040200.0.0.tar.gz) )
  - ii. For genes with multiple probes, the probe set found to be most anti-correlated with the HiSeq rnaseq (v2) normalized gene-level data sets data is used when generating the CCTs when paired rnaseq data is available. Only solid tumor patients are used for correlations. If not available, take the mean of the probe sets. For correlation ties, use first tied probe listed in file.
5. Somatic Copy Number Alterations (SCNA) data:
  - i. Somatic copy number alteration data was obtained the from the TCGA Gistic 2.0 SCNA pipeline (e.g.  
[http://gdac.broadinstitute.org/runs/analyses\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD-TP.CopyNumberLowPass\\_Gistic2.Level\\_4.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/analyses_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD-TP.CopyNumberLowPass_Gistic2.Level_4.2015040200.0.0.tar.gz) )
  - ii. Both Non-thresholded, thresholded gene-level Gistic scores, and focal region SCNA data, reported at the gene-level, were used for CCT creation.

**b. TCGA clustering tracks**

The Broad Institute's TCGA cluster analysis separates samples into clusters and then identifies the genes significantly up or down expressed in each cluster. As shown in Figure 40, SBTs are available for the up or down significantly differentially expressed genes for each cluster identified sample cluster.



**Figure 40.** Example of cluster analysis result tracks available on the TCGA portal. When users click on a “02\_clusters” node in the tree browser (red box), a series of SBTs (e.g. green box) will be available, each corresponding to the up or down differentially expressed genes from one of the identified sample cluster.

The consensus non-negative matrix factorization clustering (CNMF) and consensus ward linkage hierarchical clustering (Consensus Plus) were used by the Broad Institute in their analysis.

### c. TCGA portal TSI track data sources

TSI files containing clinical annotation data are generated for each track in the TCGA portal using the “Clinical\_Pick\_Tier1” clinical data available from the Firehose data portal. For example, the clinical data for the BRCA tissue from the 8/21/2015 TCGA result set would have been obtained from:

[http://gdac.broadinstitute.org/runs/stddata\\_\\_2015\\_08\\_21/data/BRCA/20150821/gdac.broadinstitute.org\\_BRCA.Clinical\\_Pick\\_Tier1.Level\\_4.2015082100.1.0.tar.gz](http://gdac.broadinstitute.org/runs/stddata__2015_08_21/data/BRCA/20150821/gdac.broadinstitute.org_BRCA.Clinical_Pick_Tier1.Level_4.2015082100.1.0.tar.gz)

The TSI files were additionally filtered out as follows:

- If the data for a given annotation feature is uniform.
- If the data is binary or categorical and lacks at least 5 values for EACH category.
- >90% of values for the feature are NA.

### d. Statistical correlation results

Statistical association tests (see section II.6.c) were conducted for all of the CCTs/CBTs in the portal and separate SCTs were generated for the resulting p-values, q-values, and test statistics (see Figure 41). An SBT of significant genes with a q-value < 0.05 is also generated. For each CCT/CBT, gene-level association tests are conducted for each subtrack annotation feature (see

sections II.3.c.iii and III.3.c). While all subtrack annotation features for a given track are statistically associated with the CCT/CBT data, tracks are generated and available in the “correlations” subfolder for a given subtrack annotation feature *only if at least one gene has a q-value < 0.05*. Only primary solid tumors (samples with “\_01” at the end of the sample id) were used for the tests.

In the example given in Figure 48, only a single “clinical” feature, gender (red box), was found to have at least one significantly associated gene for the precursor AUC (unshared peptides) data from the CPTAC Colorectal proteomic profile data (available by click on “profile” in purple box), while five different “molecular subtype” features were found with at least one significant result (green box).

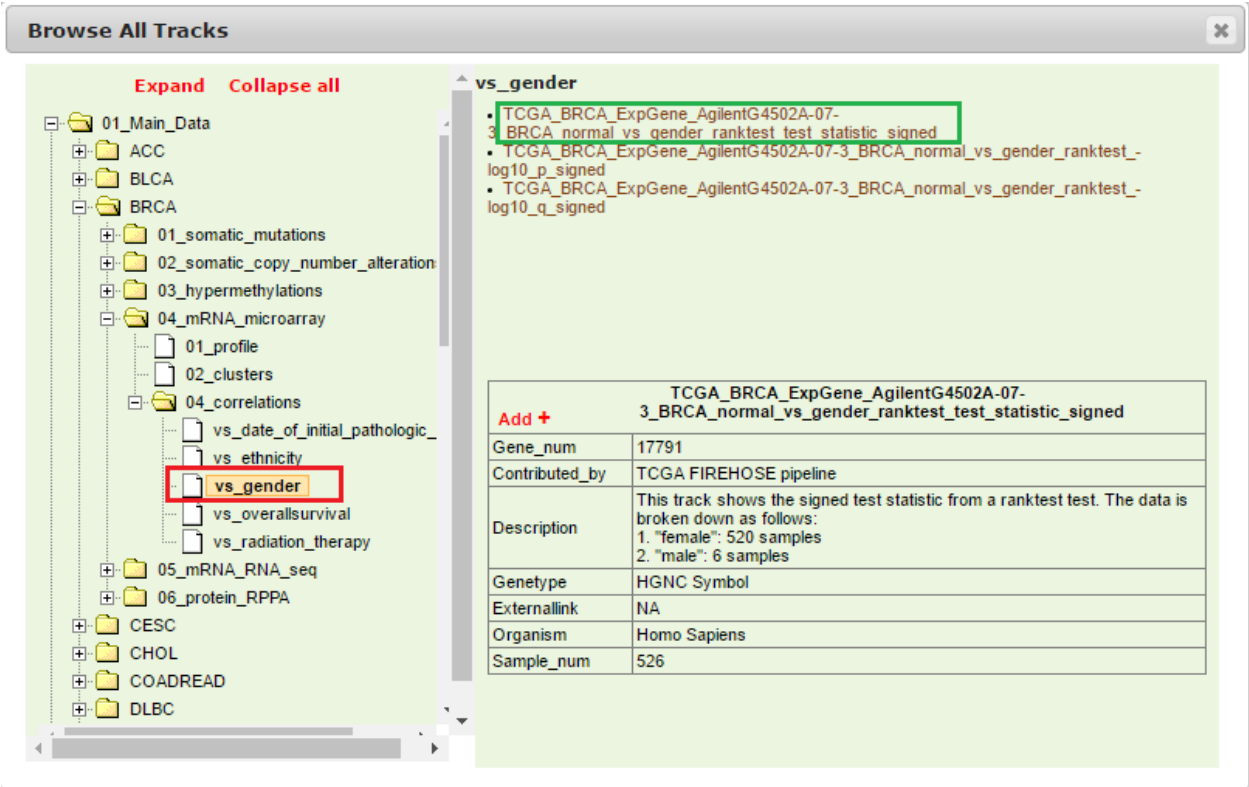


Figure 41. Example of statistical testing result tracks available on the TCGA portal. When users click on a “04\_clusters” node in the tree browser, a series of clinical attributes will be listed showing those attributes where the testing found at least one statistically significant finding (e.g. “vs\_gender” in the red box for the “gender” clinical attribute). SBTs (e.g. green box) showing the calculated test statistics, p-values, and q-values will be available for each shown test.

#### 4. Clinical Proteomic Tumor Analysis Consortium (CPTAC) Portal

As shown in Figure 42, the current version of NetGestalt CPTAC portal contains: (1) proteomic, genomic, epigenomic, transcriptomic, and clinical data for the colorectal, breast, and ovarian cohorts of both The Cancer Genome Atlas (TCGA) and CPTAC studies; (2) statistical correlation analysis results; and (3) functional tracks including Cell Map

pathways, GO Biological Processes, GO Cellular Components and GO Molecular Functions, HumanCyc pathways, KEGG pathways, NCI pathways and Reactome pathways (green box in Figure 39).

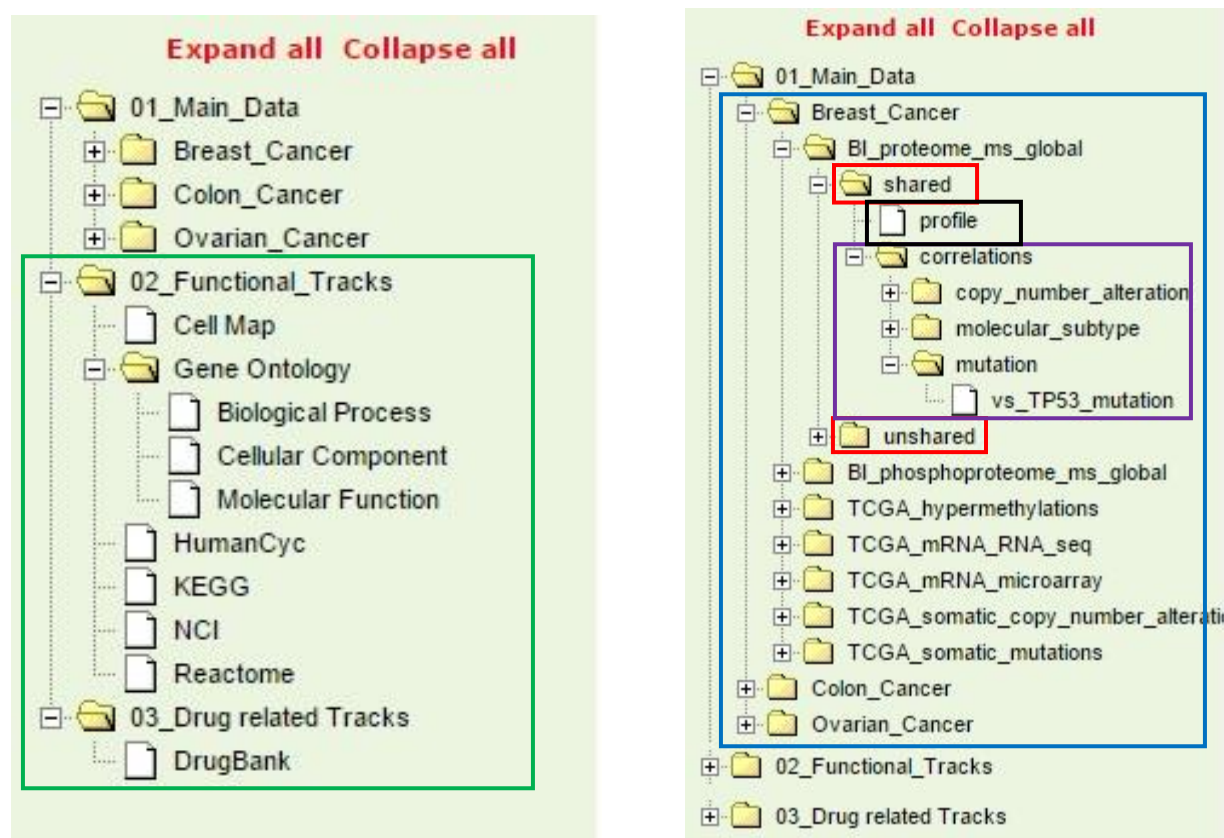
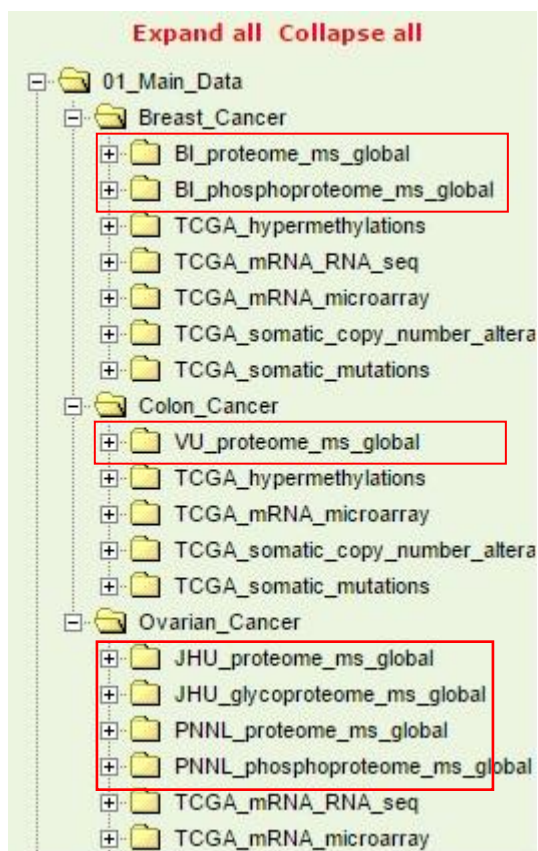


Figure 42. Overview of CPTAC and TCGA track types available. Functional tracks containing SBTs of genes associated with Gene Ontology terms or other biological databases and DrugBank target genes are highlighted in the green box. Tracks are derived from data in CPTAC and TCGA for Colorectal, breast, and ovarian cancers (blue box). Proteomic tracks from CTPAC are further divided into shared and unshared peptide data sets (red boxes). Data sets are further divided between profile tracks (clinically annotated matrices of the experimental data, shown in black box) and correlation tracks. Correlation tracks are further divided into categories (clinical annotations, copy number alterations, molecular subtypes, mutations, shown in purple box).

**a. Proteomic, phosphoproteomic, and glycoproteomic alterations from CPTAC cohorts**

Tracks derived from the CPTAC omics data for the colorectal, breast, and ovarian tumor cohorts are included for all three cancer types (red boxes in Figure 43) as follows: proteomic and phosphoproteomic breast cancer data generated by the Broad Institute; proteomic colorectal cancer data generated by Vanderbilt University; proteomic and glycoproteomic ovarian cancer data from Johns Hopkins University; and finally, proteomic and phosphoproteomic ovarian cancer data from Pacific Northwest National Laboratory.





**Figure 43. Overview of CPTAC track types available. Proteomic tracks are outlined in red.**

For the all three cancer types, the CPTAC data was obtained from the CPTAC Phase II Data Portal ([https://cptc-xfer.uis.georgetown.edu/publicData/Phase\\_II\\_Data/](https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/)). Proteomic data from CPTAC is separated into shared and unshared peptide result sets (red boxes in Figure 39A). Data sets are further divided into profile and correlation analysis results data sets.

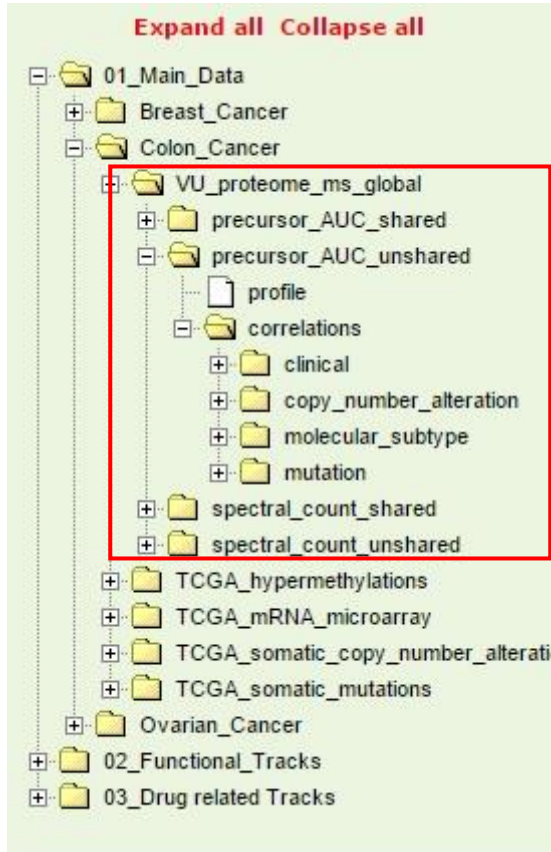
#### **i. Colorectal proteomic profile tracks**

The colorectal cancer profile tracks are CCTs containing the processed mass spectrometric proteomics data for liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based shotgun proteomic data on 90 TCGA primary solid tumor samples generated by Vanderbilt University. Data is available at the shared and unshared peptide level and reported at the protein-level (see red box in Figure 44).

The following steps were used to prepare the data tracks:

1. The “Protein Reports” containing the spectral count and precursor AUC data files were downloaded from the CTPAC data portal: [https://cptc-xfer.uis.georgetown.edu/publicData/Phase\\_II\\_Data/TCGA Colorectal Cancer/](https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/TCGA_Colorectal_Cancer/)
2. The data is normalized as follows for both spectral count and precursor AUC data at the shared and unshared peptide level:
  - a. If duplicate samples exists, the sample with the highest signal is kept.
  - b. Following deduping, the data is normalized using the global normalization method:

- i. Sum total signal for each sample. Take max observed sum and generate normalization factors for each as  $\text{max\_sum}/\text{this\_sample\_sum}$ .
- ii. Multiply all values in sample by that sample's normalization factor.
- c. Log transform the normalized values as  $\log_2(\text{value} + 1)$ .



**Figure 44.** Overview of CPTAC track colorectal data sets available (outlined in red. Precursor AUC and spectral count data from liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based shotgun proteomic data on 90 TCGA tumor samples.

## ii. Ovarian and Breast Proteomic profile tracks

The ovarian and breast cancer profile tracks are CCTs containing the processed mass spectrometric proteomics data for liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based shotgun proteomic and phosphoproteomic data on 105 TCGA primary solid breast cancer tumor samples generated by the Broad Institute and 122 TCGA primary solid ovarian cancer tumor samples generated by Pacific Northwest National Lab. Additionally, LC-MS/MS shotgun proteomic and glycoproteomic data on 84 TCGA primary solid ovarian cancer tumor samples generated by the Johns Hopkins University was also made into profile tracks (see red boxes in Figure 45). Data is available at the shared and unshared peptide level and reported at the protein-level.

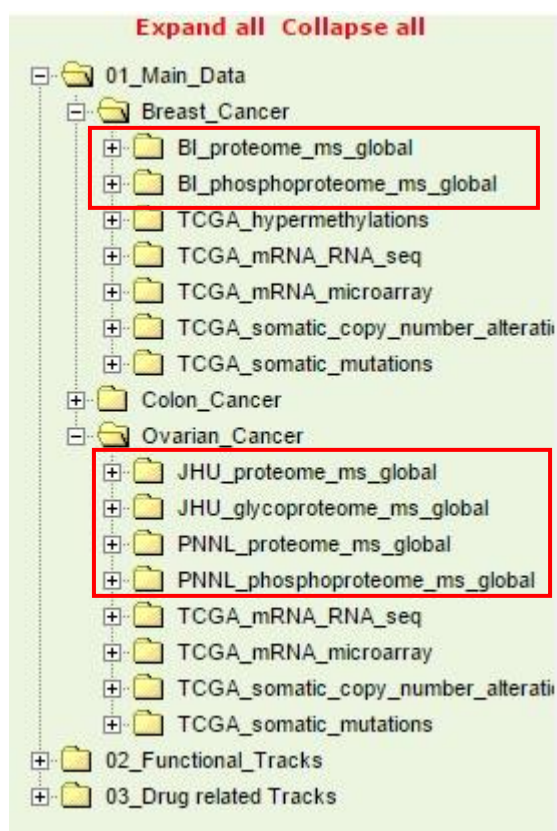
The following steps were used to prepare the data tracks:

1. The “Protein Reports” containing the iTRAQ data files were downloaded from the CTPAC data portal: <https://cptc->



[xfer.uis.georgetown.edu/publicData/Phase\\_II\\_Data/TCGA\\_Ovarian\\_Cancer/](https://xfer.uis.georgetown.edu/publicData/Phase_II_Data/TCGA_Ovarian_Cancer/)  
and  
[https://cptc-xfer.uis.georgetown.edu/publicData/Phase\\_II\\_Data/TCGA\\_Breast\\_Cancer/](https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/TCGA_Breast_Cancer/)

2. The iTRAQ data provided by CPTAC is processed as follows:
  - a. If duplicate samples exist, the sample with the highest signal is kept.
  - b. The iTRAQ data provided by CPTAC is already normalized so no additional normalization steps are required. See CPTAC normalization steps described here:  
[https://cptacdcc.georgetown.edu/cptac/documents/CDAP\\_ProteinReports\\_description\\_20140708.pdf](https://cptacdcc.georgetown.edu/cptac/documents/CDAP_ProteinReports_description_20140708.pdf)



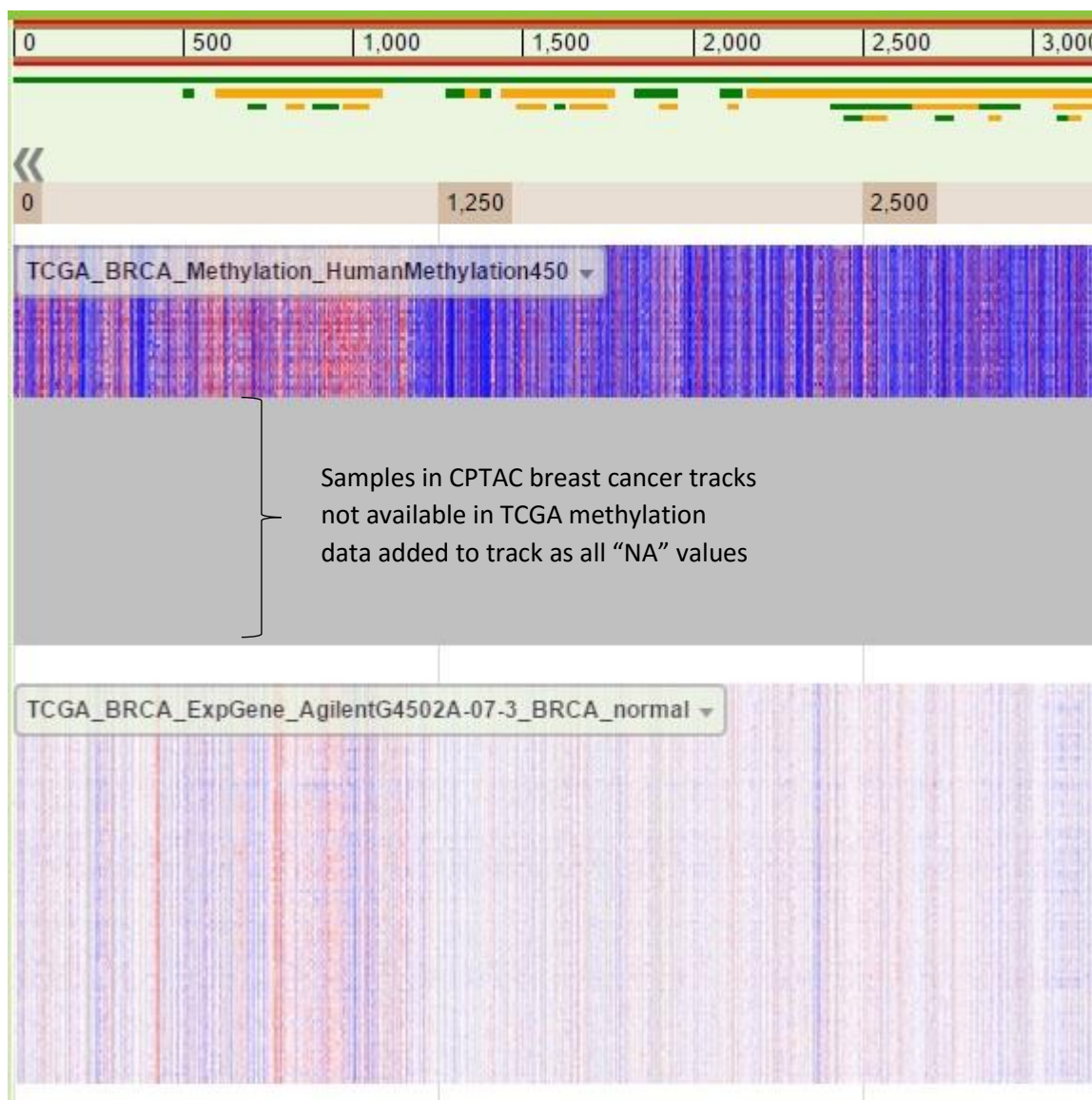
**Figure 45.** Overview of CPTAC track breast and ovarian data sets available (outlined in red). iTRAQ data from liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based shotgun proteomic data on 105 breast cancer samples generated by the Broad Institute, 122 ovarian from Pacific Northwest National Labs, and 84 samples from Johns Hopkins University.

### iii. TCGA profile tracks in CPTAC portal

The TCGA profile tracks available in the TCGA data portal (<https://gdc-portal.nci.nih.gov/>) are also available in the CPTAC portal, but are limited to ONLY those samples available in the CPTAC data sets. If samples in the CTPAC data sets are not available in the TCGA data sets, those samples will be added to the track with “NA” values for all genes (shown as a grey vertical line, see Figure 46). In the current version of the CTPAC portal, the TCGA data is taken from the 08/21/2015 version of the TCGA result sets.

The following data sets from TCGA were added to the CTPAC portal:

1. Somatic mutation data:
  - i. Mutation data was obtained from the MutSig analysis results data sets from TCGA (e.g.  
[http://gdac.broadinstitute.org/runs/analyses\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD-TP.MutSigNozzleReport2CV.Level\\_4.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/analyses_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD-TP.MutSigNozzleReport2CV.Level_4.2015040200.0.0.tar.gz))
  - ii. Silent or non-somatic mutations are filtered out.
  - iii. CBTs are generated for each version of MutSig available (version 1.5, 2.0, CV, and 2CV), where 1 = mutation present, 0 = mutation not present.
2. mRNA transcriptome microarray data:
  - i. Lowess-normalized log2-transformed transcriptome data was obtained from TCGA (e.g.  
[http://gdac.broadinstitute.org/runs/stddata\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD.Merge\\_transcriptome\\_agilent4502a\\_07\\_3\\_unc\\_edu\\_Level\\_3\\_unc\\_lowess\\_normalization\\_gene\\_level\\_data.Level\\_3.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/stddata_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD.Merge_transcriptome_agilent4502a_07_3_unc_edu_Level_3_unc_lowess_normalization_gene_level_data.Level_3.2015040200.0.0.tar.gz) )
3. RNAseq data:
  - i. Normalized RSEM gene-level RNAseq data was obtained from the HiSeq rnaseq (v2) normalized gene-level data sets (e.g.  
[http://gdac.broadinstitute.org/runs/stddata\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD.Merge\\_rnaseqv2\\_illuminahi\\_seq\\_rnaseqv2\\_unc\\_edu\\_Level\\_3\\_RSEM\\_genes\\_normalized\\_data.Level\\_3.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/stddata_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD.Merge_rnaseqv2_illuminahi_seq_rnaseqv2_unc_edu_Level_3_RSEM_genes_normalized_data.Level_3.2015040200.0.0.tar.gz) )
  - ii. The “normalized count” values provided by TCGA are  $\log_2(\text{value} + 1)$  transformed and generated as a CCT.
4. Hypermethylation data:
  - i. Methylation data was obtained from the Human Methylation 450 TCGA data set (e.g.  
[http://gdac.broadinstitute.org/runs/stddata\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD.Merge\\_methylation\\_human\\_methylation450\\_jhu usc\\_edu\\_Level\\_3\\_within\\_bioassay\\_data\\_set\\_function\\_data.Level\\_3.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/stddata_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD.Merge_methylation_human_methylation450_jhu usc_edu_Level_3_within_bioassay_data_set_function_data.Level_3.2015040200.0.0.tar.gz) )
  - ii. For genes with multiple probes, the probe set found to be most anti-correlated with the HiSeq rnaseq (v2) normalized gene-level data sets data is used when generating the CCTs when paired rnaseq data is available. Only solid tumor patients are used for correlations. If not available, take the mean of the probe sets. For correlation ties, use first tied probe listed in file.
5. Somatic Copy Number Alterations (SCNA) data:
  - i. Somatic copy number alteration data was obtained the from the TCGA Gistic 2.0 SCNA pipeline (e.g.  
[http://gdac.broadinstitute.org/runs/analyses\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD-TP.CopyNumberLowPass\\_Gistic2.Level\\_4.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/analyses_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD-TP.CopyNumberLowPass_Gistic2.Level_4.2015040200.0.0.tar.gz) )
  - ii. Non-thresholded gene-level Gistic scores were used for CCT creation.



**Figure 46. Example of two TCGA-derived profile tracks in the CPTAC portal.** The top track of methylation data contains a large number of blank sample entries (greyed out vertical lines), corresponding to the CPTAC breast cancer samples that are not available in TCGA's methylation data. The TCGA mRNA transcriptomic data (bottom track) has data for all CPTAC samples. Because sample information is limited to the samples used in the CPTAC study, any samples in the TCGA data that are not also in the corresponding CPTAC data are excluded from the tracks in this portal.

#### **b. CPTAC portal tsi track data sources**

TSI files containing clinical annotation data and other significant findings (significantly mutation genes and somatic copy number alterations from publications) are generated for each track in the CPTAC portal. The data sources for the TSI files data sources were as follows:

- e. Significantly mutated genes identified in the “sig\_genes.txt” result set available in the TCGA COADREAD study MutSig2CV TCGA analysis pipeline (e.g. [http://gdac.broadinstitute.org/runs/analyses\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD-TP.MutSigNozzleReport2CV.Level\\_4.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/analyses_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD-TP.MutSigNozzleReport2CV.Level_4.2015040200.0.0.tar.gz) )
- f. Significantly amplified or deleted focal regions identified in the TCGA COADREAD study. Focal region-level results, the significantly amplified or deleted genes are identified from the “all\_lesions.conf\_99.txt” analysis result sets available in from the TCGA “Copy Number Gistic 2.0” analysis pipeline (e.g. [http://gdac.broadinstitute.org/runs/analyses\\_2015\\_04\\_02/data/COADREAD/20150402/gdac.broadinstitute.org\\_COADREAD-TP.CopyNumberLowPass\\_Gistic2.Level\\_4.2015040200.0.0.tar.gz](http://gdac.broadinstitute.org/runs/analyses_2015_04_02/data/COADREAD/20150402/gdac.broadinstitute.org_COADREAD-TP.CopyNumberLowPass_Gistic2.Level_4.2015040200.0.0.tar.gz)). The non-thresholded values for each significantly amplified or deleted focal region are reported in the tsi file.
- g. Selected TCGA clinical data provided by the CTPAC portal:
  - d. Colorectal cancer:
    - [https://cptc-xfer.uis.georgetown.edu/publicData/Phase\\_II\\_Data/TCGA\\_Colorectal\\_Cancer/CPTAC\\_TCGA\\_Colorectal\\_Cancer\\_Protocols\\_and\\_Clinical\\_Data/COAD\\_All\\_clinical\\_features\\_TCGAbiotab\\_release1\\_090413.xlsx](https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/TCGA_Colorectal_Cancer/CPTAC_TCGA_Colorectal_Cancer_Protocols_and_Clinical_Data/COAD_All_clinical_features_TCGAbiotab_release1_090413.xlsx)
    - and
    - [https://cptc-xfer.uis.georgetown.edu/publicData/Phase\\_II\\_Data/TCGA\\_Colorectal\\_Cancer/CPTAC\\_TCGA\\_Colorectal\\_Cancer\\_Protocols\\_and\\_Clinical\\_Data/READ\\_All\\_clinical\\_features\\_TCGAbiotab\\_release1\\_090413.xlsx](https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/TCGA_Colorectal_Cancer/CPTAC_TCGA_Colorectal_Cancer_Protocols_and_Clinical_Data/READ_All_clinical_features_TCGAbiotab_release1_090413.xlsx)
  - e. Breast cancer:
    - [https://cptc-xfer.uis.georgetown.edu/publicData/Phase\\_II\\_Data/TCGA\\_Breast\\_Cancer/TCGA\\_Breast\\_Cancer\\_Metadata/BRCA\\_All\\_clinical\\_features\\_TCGAbiotab\\_r1\\_020314.xlsx](https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/TCGA_Breast_Cancer/TCGA_Breast_Cancer_Metadata/BRCA_All_clinical_features_TCGAbiotab_r1_020314.xlsx)
  - f. Ovarian cancer:
    - [https://cptc-xfer.uis.georgetown.edu/publicData/Phase\\_II\\_Data/TCGA\\_Ovarian\\_Cancer/TCGA\\_Ovarian\\_Cancer\\_Metadata/OV\\_All\\_clinical\\_features\\_TCGAbiotab\\_CPTAC\\_S020.xlsx](https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/TCGA_Ovarian_Cancer/TCGA_Ovarian_Cancer_Metadata/OV_All_clinical_features_TCGAbiotab_CPTAC_S020.xlsx)
- h. Additional annotation sources from the following publications:
  - g. Colorectal cancer data:
    - i. CPTAC clinical data from Zhang, et al. (2014), including the proteomic subtypes identified in the study: <http://www.nature.com/nature/journal/v513/n7518/extref/nature13438-s1.xlsx>
    - ii. Clinical data from TCGA Colorectal study (2012), including the methylation clustering, MLH1\_silencing, and MSI\_status subtypes, and hypermutation status:

<http://www.nature.com/nature/journal/v487/n7407/extref/nature11252-s3.zip>

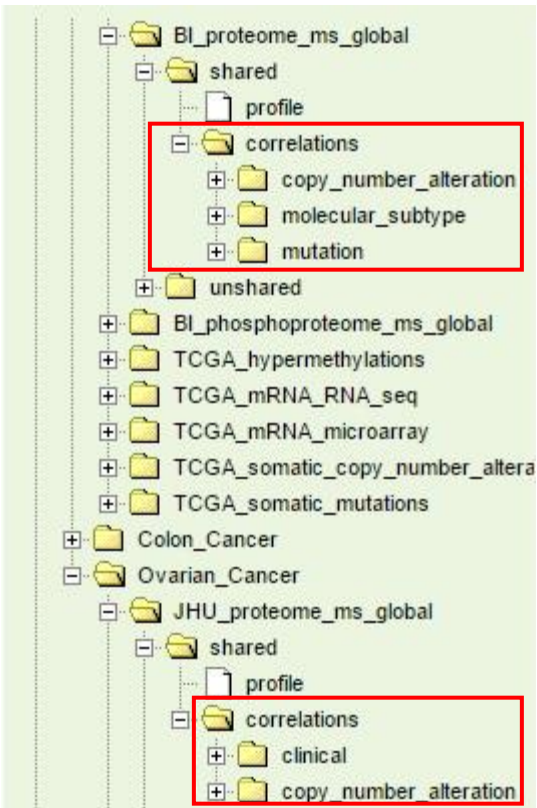
- h. Breast cancer data:
  - i. Clinical data from 2012 TCGA Breast study, including various categories of clustering:  
<http://www.nature.com/nature/journal/v490/n7418/extref/nature11412-s2.zip>
- i. Ovarian cancer data:
  - i. Clinical data from 2012 TCGA Ovarian study:  
<http://www.nature.com/nature/journal/v474/n7353/extref/nature10166-s2.zip>
- i. TSI files were additionally filtered out as follows:
  - j. If the data for a given annotation feature is uniform.
  - k. If the data is binary or categorical and lacks at least 5 values for EACH category.
  - l. >90% of values for the feature are NA.

### c. Statistical correlation results

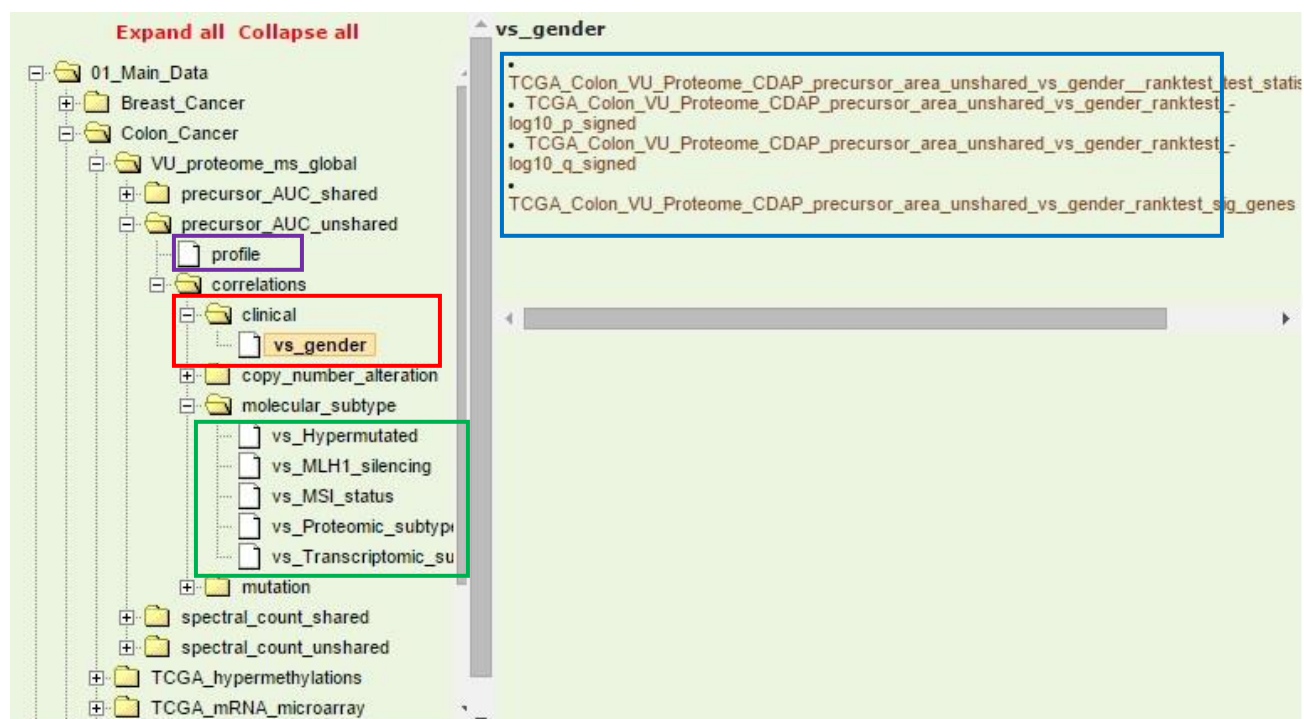
Statistical association tests (see section II.6.c) were conducted for all of the CCTs/CBTs in the portal and separate SCTs were generated for the resulting p-values, q-values, and test statistics (see Figure 47). An SBT of significant genes with a q-value < 0.05 is also generated. For each CCT/CBT, gene-level association tests are conducted for each subtrack annotation feature (see sections II.3.c.iii and III.3.c). While all subtrack annotation features for a given track are statistically associated with the CCT/CBT data, tracks are generated and available in the “correlations” subfolder for a given subtrack annotation feature *only if at least one gene has a q-value < 0.05*.

In the example given in Figure 48, only a single “clinical” feature, gender (red box), was found to have at least one significantly associated gene for the precursor AUC (unshared peptides) data from the CPTAC Colorectal proteomic profile data (available by click on “profile” in purple box), while five different “molecular subtype” features were found with at least one significant result (green box).





**Figure 47.** Example of statistical analysis track subfolders in the CPTAC portal tree browser (red boxes). Statistical association tests were conducted for all CPTAC and TCGA-derived profile tracks in the CPTAC portal.



**Figure 48.** Example of tracks available from the statistical association tests conducted on the precursor AUC (unshared peptides) CCT profile track (purple box). Each of the subtrack annotation features for the profile track were tested against each gene in the track. Features with at least one significant gene ( $q\text{-value} < 0.05$ ) have SCTs generated for the test statistic, p-values, q-values, and an SBT of significant genes. Each subtrack annotation features falls into one of four categories: clinical, copy\_number\_alteration, molecular subtype, and mutation. In the above example, a single “clinical” feature (gender) was found to be significantly associated with at least one gene (red box). The resulting tracks are shown in the panel on the right (blue box).

## References:

Cancer Genome Atlas Network (2011) Integrated genomic analyses of ovarian carcinoma, *Nature*, 474, 609-615.

Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer, *Nature*, 487, 330-337.

Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours, *Nature*, 490, 61-70.

Chartier, M., et al. (2013) Kinome Render: a stand-alone and web-accessible tool to annotate the human protein kinome tree, *PeerJ*, 1:e126, <https://doi.org/10.7717/peerj.126>.

Cheung, H.W., et al. (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer, *Proceedings of the National Academy of Sciences of the United States of America*, 108, 12372-12377.

Liberti, S, et al. (2013) HuPho: the human phosphatase portal, *The FEBS Journal*, Volume 280, Issue 2, 379-387.

Shi, Z., Wang, J. and Zhang, B. (2013) NetGestalt: integrating multidimensional omics data over biological networks, *Nat Methods*, 10, 597-598.

Subramanian, A., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15545-15550.

Wang, J., et al. (2013) Integrative genomics analysis identifies candidate drivers at 3q26-29 amplicon in squamous cell carcinoma of the lung, *Clinical cancer research: an official journal of the American Association for Cancer Research*, 19, 5580-5590.

Wingender, E., Schoeps, T. and Dönitz, J (2013) TFClass: An expandable hierarchical classification of human transcription factors, *Nucleic Acid Research*, 41, D165-D170.

Zhang, B., et al. (2014) Proteogenomic characterization of human colon and rectal cancer, *Nature*, 513, 382-387.