I. Use WebGestalt to perform pathway analysis for a gene list

1. Go to http://www.webgestalt.org
2. Read INTRODUCTION to understand the workflow
3. Read DATA SOURCE to know the pathway/gene set databases used for the enrichment analysis
4. Click on "Sample data", save "Interesting Gene List" (interestingGenes.txt) and "Reference Gene List" (referenceGenes.txt) on your computer. For future analysis of your own data, your data files should be prepared in the same format, and special characters are not allowed in the file name.
5. Click on "START" on the homepage
6. Upload interesting gene list
    a. Select_Organism_from_Drop_Down_Menu: hsapiens
    b. Select_ID_Type_from_Drop_Down_Menu: hsapiens_gene_symbol
    c. Upload gene list (click on the *i* button to get information on file format etc)
        i. Choose file: interestingGenes.txt
    d. Enter
7. Get information on the uploaded gene list
    a. Mapping information at the top
    b. Mapping table at the bottom
    c. GO Slim Classification
8. Perform GO enrichment analysis
    a. Select Enrichment analysis type: GO Analysis
    b. Upload User Reference Set File and Select ID Type
        i. Choose File: referenceGenes.txt
        ii. Select_ID_Type_from_Drop_Down_Menu: hsapiens_gene_symbol
    c. Statistical Method: Hypergeometric
    d. Multiple Test Adjustment: BH
    e. Significance Level: Top10
    f. Minimum Number of Genes for a Category: 2
    g. Run Enrichment Analysis
9. Explore enrichment analysis results
    a. View results
        i. Enrichment GO directed acyclic graphs (DAGs)
        ii. Click on "programmed cell death", one of the enriched categories
        iii. Detailed statistical results and the list of overlapping genes in the category
    b. Export TSV only
    c. Export Complete Results Package
10. Repeat GO enrichment analysis using significance level 0.01 (step 6e)
11. Perform enrichment analysis based on other gene set databases, using significance Level: top 10
    a. KEGG Analysis: click a pathway name in the output will show pathway map with the interesting genes highlighted

b. Wikipathways Analysis: click a pathway name in the output will show pathway map, gene highlighting function is not working currently
c. Transcription Factor Target Analysis: notice NFAT
d. microRNA Target Analysis: notice mir-377
e. protein interaction network analysis: notice Module_414, which is related to apoptosis
f. Disease Association Analysis: Colorectal Neoplasms ranked No.1
g. Phenotype Association Analysis: notice Neoplasm of the large intestine

II. Use NetGestalt to perform pathway and network analysis for a gene list

1. Go to http://www.netgestalt.org
2. What would you like to do: Analyze your own data
3. Select an organism: Human
4. From the popup “Upload Data” window, save “crcSignature.sbt” to your computer. For future analysis of your own data, your data files should be prepared in the same format, and end with .sbt or .sbt.txt
5. Upload data
    a. Select a data type: gene list (sbt)
    b. Choose a file: crcSignature.sbt.txt
    c. ID type: hgnc_symbol
    d. Submit
6. Default view is the chromosome_view, i.e. genes are ordered based on their locations on the chromosome. Green and yellow bars at the top represent chromosomes and chromosomal bands
7. Switch to the network_view
    a. View->select->iRef
    b. Now genes are ordered based on linearized protein-protein interaction network according to the hierarchical modular structure of the network, green and yellow bars at the top represent hierarchical network modules
8. Perform pathway enrichment analysis (steps a-i can be done in the chromosome_view as well)
    a. Mouse over the track name “CRCsignature_xxxxx” (suffix “xxxxx” is added by the system to ensure uniqueness of the track names),
    b. Click “Gene Set Enrichment”
    c. Select “cellmap” and FDR cutoff 0.05
    d. Enriched cellmap pathways are listed on the left panel
    e. Click on “Wnt”, select “All genes” to add all “wnt” genes annotated in cellmap as a new track
    f. Mouse over the track name “CRCsignature_xxxxx”
    g. Click “Gene Set Enrichment”
    h. Select “kegg” and FDR cutoff 0.05
    i. Enriched KEGG pathways are listed on the left panel
    j. Find “Wnt signaling pathway” from the enriched list, click on it, select “All genes” to add all “Wnt signaling pathway” genes annotated in KEGG as a new track. Note the inconsistency between the cellmap and KEGG annotations.
    k. To find CRC signature genes that are annotated as Wnt signaling pathway genes in both cellmap and KEGG, use the “Track Comparison” feature available on the left panel
    l. Click on the middle section in the venn diagram and provide a track title, e.g. crc_wnt, to create a new track based on the overlapping genes (Note: any

sections in the venn diagram can be used to create new tracks, you may also click on the white space to add the union as a new track.)

m. Mouse over the name of the new track, “crc_wnt” in this case, click on “Node-link graph (Present Nodes)” to retrieve a network graph for genes in this track. Network graph is interactive and can be exported.
n. Close the “Graph View” and then mouse over the three wnt tracks to delet these tracks.

9. Perform network analysis
    a. Network retrieval
        i. Mouse over the track name “CRCsignature_xxxxx”
        ii. Click on “Node-link graph (Present Nodes)” to retrieve a network graph for genes in this track.
        iii. Note the graph is too busy to be easily understood
    b. Network module enrichment analysis
        i. Mouse over the track name “CRCsignature_xxxxx”
        ii. Click “Network analysis”
        iii. Under the “Module enrichment” section, select “Identify enriched network modules”, and select FDR cutoff 0.05
        iv. Four enriched modules are identified and listed on the left panel
        v. Click “Add all related modules” and provide a track title, e.g. all enriched modules, to add all enriched modules as a new track
        vi. Click on a green or yellow bar (in the top panel) corresponding to one of the enriched modules, e.g. level4_module28 to zoom into this network region. Note gene names appear after zooming
        vii. Mouse over the track name “CRCsignature_xxxxx”
        viii. Click on “Node-link graph (Visible Range)” to retrieve a network graph for all genes in the visible range, with genes in the track highlighted in green. This represents the enriched module with the uploaded crc signature genes in the module highlighted.
        ix. Alternatively, click on “Node-link graph (Present Nodes)” will create a network graph for genes in the visible range and also present in the track. This represents the network of crc signature genes in this module.
        x. Click on the top green bar to zoom out to the whole network, and similar graphs can be generated for all other enriched modules. Note if the visible range is too big, only “Node-link graph (Present Nodes)” is available.
        xi. Mouse over the “all enriched modules” track and delete it.
    c. Network expansion (to identify other genes that might be related to CRC based on their enriched connectivity to the seed genes)
        i. Click on the top green bar to zoom out to the whole network
        ii. Mouse over the track name “CRCsignature_xxxxx”
        iii. Click “Network analysis”

iv. Under the "Network expansion" section, select "All enriched neighbors (including seeds)", select FDR cutoff 0.05, and enter a new track name, e.g. "crc signature expanded", and then click on go

v. Use the "Track Comparison" feature on the left panel to check the number of new genes identified (9 in this case)

vi. Mouse over the track name "crc signature expanded", click on "Node-link graph (Present Nodes)" will create a network graph for both the seed genes (red) and the added genes (blue). Note the added genes include TWIST1, SRC, among others of potential interest

vii. Mouse over the "crc signature expanded" track and delete it

d. Network prioritization (to identify the most important seed genes based on their enriched connectivity to other seed genes)

i. Click on the top green bar to zoom out to the whole network

ii. Mouse over the track name "CRCsignature_xxxxx"

iii. Click "Network analysis"

iv. Under the "Gene prioritization" section, select "Enriched seeds", select FDR cutoff 0.05, and enter a new track name, e.g. "crc signature prioritized", and then click on go

v. Mouse over the track name "crc signature prioritized", click on "Node-link graph (Present Nodes)" will create a network graph for the prioritized seed genes, which include TP53, KRAS, among others.

vi. To visualize these genes in the context of other crc signature genes, use the "Track Comparison" feature on the left panel, click on the white space of the resulted venn diagram and provide a track name, e.g. "crc signature prioritized 1" will create a new track with the union of genes in both track. Notably, classification of genes from different sections of the venn diagram is carried over to the new track. Therefore, the "Node-link graph (Present Nodes)" created from this track will highlight the prioritized seed genes in a different color.

III. Use NetGestalt to perform differential expression and pathway/network analysis for a gene/protein expression data set (matrix)

1. Go to http://www.netgestalt.org
2. What would you like to do: Analyze your own data
3. Select an organism: Human
4. From the popup "Upload Data" window, save "crc_tcga_90_proteomics.cnt" and "crc_tcga_90_proteomics.tsi" to your computer. The cnt file stores the gene by sample spectral count data from a colorectal cancer proteomics study (Zhang et al. Nature 2014). The tsi file stores sample annotation information on their hypermutation status. In this case, there are 90 CRC tumors, including 18 hypermutated, 67 non-hypermutated, and 5 unknown samples. We will use NetGestalt to perform differential expression analysis for this data set and then to perform pathway and network analysis. For future analysis of your own data, your data files should be prepared in the same format. Sample annotation files should be prepared in the tsi format, and file name should end with .tsi or .tsi.txt. If the expression data is count data (e.g., RNA-Seq raw counts or spectral counts from proteomics), it should be prepared as cnt files (file name end with .cnt or .cnt.txt). If the expression data is normalized continuous data (e.g., RSEM normalized RNA-Seq data, RMA normalized microarray data, normalized proteomics label free intensity data, or normalized iTRAQ data, etc), it should be prepared as cct files (file name end with .cct or .cct.txt). cnt and cct files are in the same format, but one is for count data (which will be analyzed by voom/limma) and the other is for continuous data (which will be analyzed by limma).
5. Upload data
   a. Select a data type: data matrix – integer (cnt)
   b. Choose a file: crc_tcga_90_proteomics.cnt.txt
   c. Choose a sample information file: crc_tcga_90_proteomics.tsi.txt
   d. ID type: hgnc_symbol
   e. Submit
6. Default view is the sorted_view, in which genes are ordered based on their direction and significance of differential expression (i.e., signed –logp values). Four tracks are presented in the sorted_view, included the normalized spectral count data track (note that genes for which the number of sample with 1 count per million < n (n is the number of samples in the smaller group for the group comparison) is considered not quantifiable and thus removed), signed_minus_log_p, signed_minus_log_adj_p, and logFC. Clicking the "double arrow" button located above the top-left of the first track can hide these track names. Dragging a track name can change the vertical position of the track.
7. Improve heatmap visualization
   a. Mouse over the track name "crc_tcga_90_proteomics_cnt_ xxxxx"
   b. Click on "Subtrack Annotation", click on the "+" sign in the left panel, then select "Hypermutated" from the dropdown menu. Samples are now sorted based on their hypermutation status

   c. Close the sample annotation heatmap by click on the “x” sign

8. Identify genes with FDR<0.01 and absolute log fold change >1
   a. Mouse over the track name “Signed_minus_log_adj_p_xxxxx”
   b. Click on “Value based filtering”
   c. In the pop-up window, select less than -2 or greater than 2, and enter a title for the new track, e.g. “FDR001”
   d. Mouse over the track name “logFC _xxxxx”
   e. Click on “Value based filtering”
   f. In the pop-up window, select less than -1 or greater than 1, and enter a title for the new track, e.g. “2fold”
   g. Using the “Track Comparison” feature in the left panel to get the overlapping genes between “2fold” and “FDR001”, i.e., click on the intersection in the venn diagram and enter a track title, e.g. “2fold_and_FDR001”
   h. Mouse over the track name “2fold_and_FDR001”
   i. Click on “Presence-based filtering”
   j. In a new window, the current view is filtered by genes in the “2fold_and_FDR001” track. Click on “Show sample heatmap” at the top, and the sample heatmap will be shown
   k. Use mouse to move to the left or right to see all gene names and corresponding data in the heatmap
   l. Close the new window
   m. The “2fold_and_FDR001” track can be used for pathway and network analysis as described in Section II
   n. Delete the “2fold_and_FDR001”, “2fold”, and “FDR001” tracks.
9. Perform pathway enrichment analysis based on the signed minus logp values
   a. Mouse over the track name “Signed_minus_log_p_xxxxx”
   b. Click on “Gene Set Enrichment”
   c. In the popup window, select “kegg” and FDR cutoff 0.05
   d. Enriched pathways are identified using GSEA and shown in the left panel
   e. From the list of enriched pathways, add the first pathway as a new track by clicking on the pathway name, and then select “All”, and then Add. (The overlapping/leading edge option will only add leading edge genes).
   f. Click on the second pathway name, and then select “All, and then Add
   g. Other pathways can be added similarly
   h. Delete all added pathway tracks
10. Perform network analysis based on the signed minus logp values
   a. From the “View” menu at the top, select the iRef view
   b. Improve heatmap visualization as described in step 7
   c. Mouse over the track name “Signed_minus_log_p_xxxxx”
   d. Click on “Network Analysis”
   e. In the popup window, select FDR cutoff 0.1, Click “GO”
   f. Enriched network modules are identified using GSEA and shown in the left panel

g. Click on “iRef_Level3_Module1_N”, and then select the overlapping/leading edge option to add leading edge genes in the module as a new track, enter a track title, e.g. “l3m1_leading”
h. Click on the green bar (in the top panel) corresponding to Level3_Module1 to zoom into this network region.
i. Use the “Track Co-visualization” feature in the left panel to generate a network diagram for this module, select “l3m1_leading” for node color and “Signed_minus_log_p_xxxxx” for node border color, then click on the “G” button
j. A node-link diagram is available in the pop-up window. Node positions can be manually rearranged, and the graph can be exported.
k. Close the “Graph View”
l. Click on the top green bar to zoom out to the whole network
m. Delete “l3m1_leading”
n. Other modules can be explored individually following steps g-m
o. Alternatively, click “Add all related modules” in the left panel and provide a track title, e.g. all enriched modules, to add all enriched modules as a new track

IV. Use Gene2Net to perform network diffusion analysis for a gene list

1. Go to http://www.webgestalt.org
2. Click on “Sample data”, click on “Interesting Gene List”, and copy the list of genes to your clipboard
3. Network expansion
    a. Go to http://www.gene2net.org
    b. Under select portal, select PPI
    c. Click on “START”
    d. Select data set: Human PPI
    e. Select network construction method: Network_Expansion
    f. Select number of top ranking neighbors: 10
    g. Select significant level method for enrichment analysis: FDR
    h. Select FDR threshold: 0.05
    i. Highlight: Neighbors
    j. Input gene symbols: paste the gene list on your clipboard
    k. Submit
4. Network prioritization
    a. Go to http://www.gene2net.org
    b. Under select portal, select PPI
    c. Click on “START”
    d. Select data set: Human PPI
    e. Select network construction method: Network_Retrieval_Prioritization
    f. Select number of top ranking neighbors: 10
    g. Select significant level method for enrichment analysis: FDR
    h. Select FDR threshold: 0.05
    i. Input gene symbols: paste the gene list on your clipboard
    j. Submit