GCBA/MGCB/BMI 815, Fall 2017 10/27/17 Instructor: Babu Guda

Due date: 1am, 11/3/17 (Friday)

Midterm-2 Project Questions: Total 30 points

Question-1 (7 points)

Using the UCSC Genome Browser:

1. Based on the "Gene Expression in 53 tissues track", list the tissue in which this gene exhibits the highest median expression.
2. Based on the "Multiz Alignments of 100 vertebrates" comment on the level of conservation of this gene for the coding exons. Note any aberrations or lessening of conservation across the eight species listed. Note: a yellow bar in this track indicates that genome is incomplete at that location.
3. Based on the "Single Nucleotide Polymorphisms" (SNP) track, list the number of non-synonymous + splicing variants present and their dbSNP IDs (e.g., rs123456789).
4. Connect to the Cancer Genome Tracks from the 'track hubs' link from UCSC browser. Search your gene from the 'Search Term' window. Based on the 'TCGA RNA-seq average expression per tumor type' track, list at least one cancer type in which this gene is downregulated (if any exist), and also at least one in which this gene is upregulated (if any exist). In none exist, report the same.

Using IGV:

5. Download IGV (https://goo.gl/PBamc1) and the alignment files (https://goo.gl/hX4XGk). After loading the alignment file (the one ending in ".bam") into IGV, navigate to your gene and, using the gray histogram next to "exam.bam Coverage", count all of the SNPs **in the coding exons** and state how many there are. Then include a screenshot of one of them to paste into your Word document. There might be only one for your gene.

Question-2 (8 points)

Perform the differential gene expression analysis on your assigned GSE dataset and fold-change values using ExAtlas web site (https://lgsun.irp.nia.nih.gov/exatlas). Note that all datasets use Mouse as the organism. Present the following results (copy and paste the images) in your report and explain your observations on the following results in 1-2 sentences each.

(i) Gene expression heatmap (click on 'Make heatmap' button)
(ii) Two PCA figures (first select 'Show replications' and 'PC gene clusters' before clicking the PCA button)
(iii) Scatter plot figure from 'Pairwise Comparison'
(iv) Perform geneset enrichment analysis using Gene Ontology and present the 'z_value_combined' plot and the hierarchical clustering heatmap.

Question-3 (5 points)

Perform gene enrichment analysis on the gene list assigned to you using WebGestalt (http://www.webgestalt.org/option.php). Choose the following options: 'Homo sapiens' as the 'organism of interest'; 'genesymbol' for 'Select Gene ID Type'; and 'genome_protein-coding' (if needed) for the 'Select Reference set for Enrichment Analysis' options. The rest of the options are assigned to you in the spreadsheet, and any unassigned parameters should use the default values.

Please submit only the following by pasting the table or image (as relevant) in your word document:

a. No. of total genes in your genelist, numbers mapped and unmapped?
b. If you are assigned with 'ORA analysis', submit the 'Top 10 hits' Table (can be found in the 'Enrichment Results' tab on the left)
c. If you are assigned with 'NTA analysis', submit the 'Sub-network graph' image. The image can be saved using 'Export as' option.

Question-4 (5 points)

Perform gene enrichment analysis on the gene list assigned to you using DAVID (https://david.ncifcrf.gov/). On the 'Upload' tab, paste your gene list (step 1), select 'OFFICIAL_GENE_SYMBOL' (step 2) and 'Gene list' (step 3). On the 'List' tab, select 'Homo sapiens' as the species of choice, select the 'list' in the List Manager and click 'Use' to start the analysis. Analyze the gene list for enrichment in following biological contexts. All results should be copied and pasted into your report.

i. Disease Category: Select OMIM_DISEASE category and click on the horizontal 'Blue bar' to see the list of diseases. Report the top 5 diseases from this table. If you get less than 5 or none, report the same.
ii. Gene Ontology: Select GOTERM_BP_DIRECT category and click on the 'Chart' button to see the list of Biological Processes. Report the top 5 processes from this table. If you get less than 5 or none, report the same.
iii. Pathway Category: Select KEGG_PATHWAY category and click on the 'Chart' button to see the list of KEGG pathways. Report the top 5 pathways from this table. If you get less than 5 or none, report the same.
iv. Protein Domains: Select INTERPRO category and click on the 'Chart' button to see the list of InerPro protein domains. Report the top 5 domains from this table. If you get less than 5 or none, report the same.

Question-5 (5 points)

Download the software (http://software.broadinstitute.org/gsea/downloads.jsp) and GSEA datasets assigned to you (http://software.broadinstitute.org/gsea/datasets.jsp) Perform GSEA analysis and the leading edge analysis using the assigned datasets and

MSigDB genesets and submit the following in your report. Use the default parameters unless specified.

(i) Name of GSEA dataset and phenotype labels (1 sentence)
(ii) Name of MSigDB geneset you are using and its characteristics (1 sentence)
(iii) The GSEA analysis main report (copy and paste)
(iv) Open the GSEA report and answer the following questions:
   a. Open the snapshot of the enriched gene sets in the first phenotype. Click on the first enriched gene set and paste the plot below. Interpret the plot briefly, describing the Enrichment Score (ES), Normalized ES (NES), p-value, FDR value, and the portion of the ranked gene list contributing to the ES.
(v) Select top 10 enriched gene sets and perform leading edge analysis, as described in the class. Copy and paste all four plots generated by the leading-edge analysis and interpret each plot briefly.

How to submit your project?

- Your final project should contain a single Word file containing your responses to all questions.
- All the results (tables, screenshots or other images) should be inserted into your Word document. The Word file name MUST start with your last name.
- Your description of results should be brief and to the point (1-2 sentences)
- Please email your project file to babu.guda@unmc.edu by 1am on 11/3/17 (Friday)