Deepknomics – Interns problem statement

Problem statement –

- Download RNA expression data of a cancer type of your choice from TCGA.
- Match the samples with relevant tissue expression data from GTEX.
- Normalize the expression across TCGA and GTEX.
- Generate plots to show expression difference of a gene of your choice.
- Submit the code and a summary of your findings.

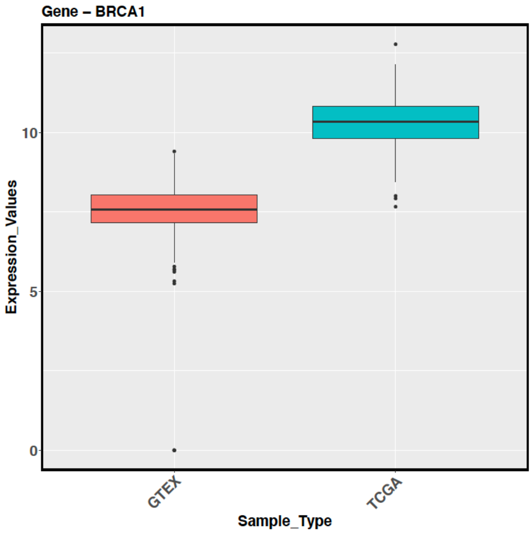
Solution –

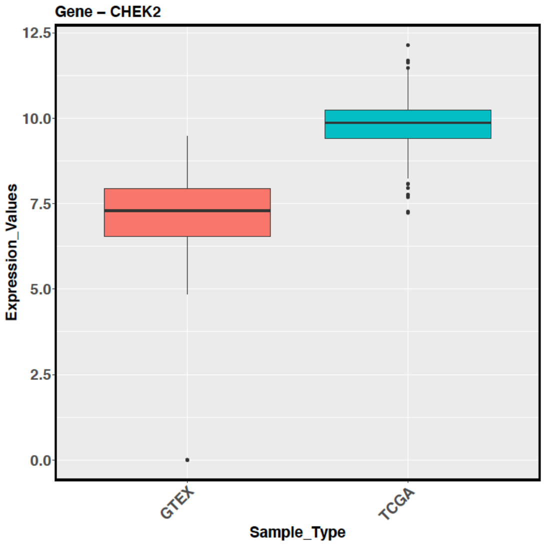
- **Data collection**
    - Download the data from TOIL TCGA GTEx cohort RNAseq of expected count data from Xena hubs which has all data related to TCGA and Gtex.
    - https://toil.xenahubs.net/download/gtex_gene_expected_count.gz - and unzipped it in the working directory
    - https://toil.xenahubs.net/download/tcga_gene_expected_count.gz - and unzipped it in the working directory
    - Phenotypic information of all samples from TCGA abd GTEX data –
        - It has mapping of samples associated to TCGA and GTEX as per the tissue and disease.
        - Fetched Primary tumor and tissue as colon cancer for TCGA cancer data
        - Fetched GTEX and Colon as a tissue for GTEX normal data.
        - https://toil.xenahubs.net/download/TcgaTargetGTEX_phenotype.txt.gz and unzipped it in the working directory
        - Ensembl gene id and gene name mapping – File is downloaded from – https://toil-xena-hub.s3.us-east-1.amazonaws.com/download/probeMap%2Fgencode.v23.annotation.gene.probemap

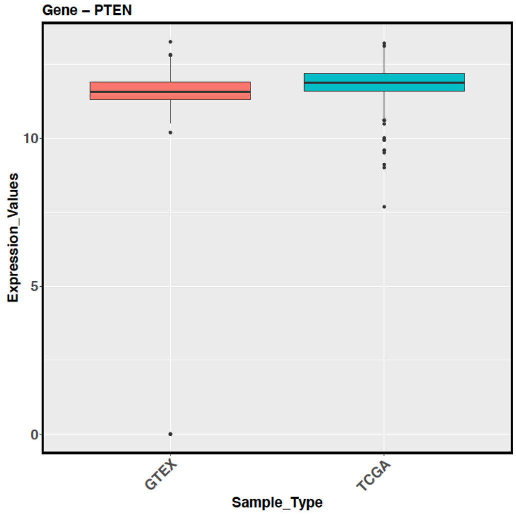
- **Data processing and normalization**
    - Data is fetched for tissue of interest.
        - From TCGA data (expected count) – Primary Tumor – Colon Cancer
        - From GTex data (expected count) – GTEX – Normal Tissue - Colon
    - Normalization is done after data is combined – log2+1
    - Could also try with TPM and other count metrics.

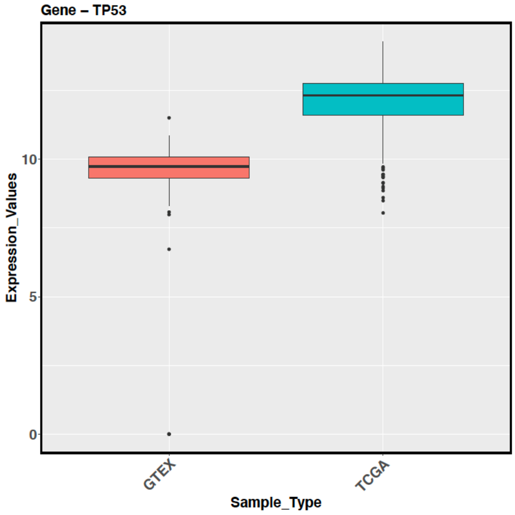
- **Differential expression/comparison analysis for gene of interest and plotting using ggplot2 as a boxplot.**

- **Summary**
    - Its observed that cancer causing genes are highly expressed in TCGA – Colon Tumor samples as compared to GTEX which are normal tissue of colon for the gene of interest selected.
    - Data processing, transformation, normalization and plotting is done in R language.
    - Boxplot is plotted for the gene of interest using ggplot2 in R.
    - Please see the boxplot for some of the genes as a example from the script written.
    - The genes which are highly expressed in TCGA than GTex, - this paper is cross verify our results - **GEPIA**: a web server for cancer and normal gene expression profiling and interactive analyses.









References

- https://xenabrowser.net/datapages/
- https://www.nature.com/articles/s41467-017-01027-z
- https://rpubs.com/woodhaha/380331
- http://gepia.cancer-pku.cn/index.html - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570223/
- https://www.sciencedirect.com/science/article/pii/S266616672200048X#tbl1