R Biostatistics Course

The HMS Research Computing R Biostatistics course is a three class, advanced instructional course covering the basics of RNA-seq analysis with Bioconductor and the R statistical programming language. Bioconductor provides tools for the analysis and comprehension of all types of high-throughput genomic data. Students should have a beginner's level of proficiency in R programming and an understanding of basic statistical principles before registering for the course. This course is also taught as part of the 20.440 Biological Networks course at MIT.

The course covers standard supervised statistical approaches for the comprehensive analysis of a published Cancer Genome Atlas (TCGA) human breast cancer RNA-seq dataset. Topics include edgeR differential gene expression analysis and GOSeq functional enrichment analyses of gene ontology terms and KEGG pathways. Data visualization techniques are emphasized, and each two-hour class includes a lecture and R practicum. Course registration includes all 3 classes.

Comprehensively commented R scripts are provided to the student, as the objective of the course is to learn common biostatistical methods used for RNA-Seq analysis. Students are strongly encouraged to use personal laptops for the course. Students are also encouraged to download and install the latest version of R prior to the first class. The latest version of R can be found in The Comprehensive R Archive Network, http://cran.r-project.org/

Due to potential compatibility issues with R studio, we suggest students download and install the latest version of Sublime Text, a code text editor for both (Mac) OS X and Windows, https://www.sublimetext.com/3

Supervised Differential Gene Expression and Functional Enrichment Analyses of a Published TCGA Human Breast Cancer Dataset

- Differential Gene Expression Analysis with edgeR
- Functional Enrichment of Gene Ontology Terms with GOSeq Analysis
- Functional Enrichment and Visualization of Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways with GOSeq Analysis and Pathview, a pathway-based data integration and visualization tool

---

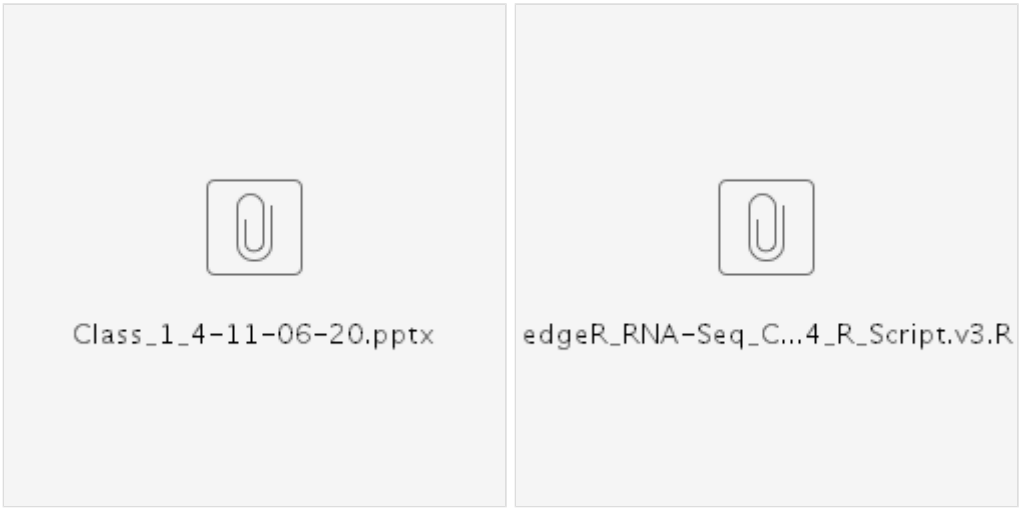
Class Files and Example Scripts

Class files and example R scripts for HMS Research Computing's three-class R/Biostatistics Course with Dr. Tom Chittenden are posted two days prior to the start of each class. Kristina Holton of HMS Research Computing serves as a teaching assistant for the course

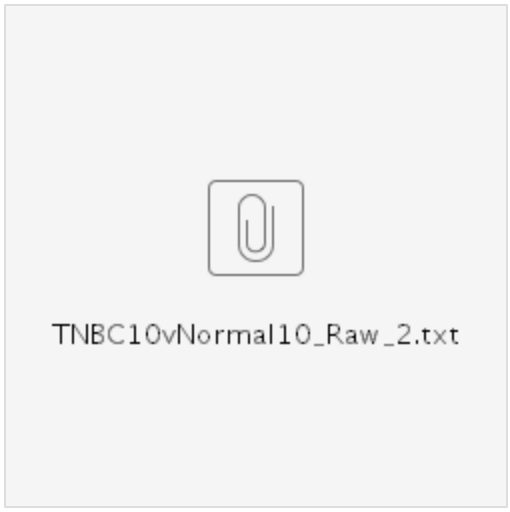
Biostatistics Class 1 Files and Example R Scripts

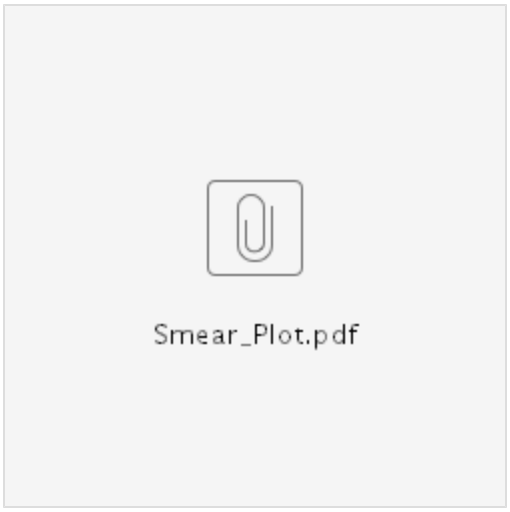
Classes 1-3 highlight specific methods associated with Differential Gene Expression and Functional Enrichment Analyses of a Published TCGA Human Breast Cancer RNA-seq Dataset. Class 1 covers Empirical analysis of digital gene expression data in R with the edgeR package by Robinson et al, 2010 at the Bioinformatics Division of the Walter and Eliza Hall Institute of Medical Research.

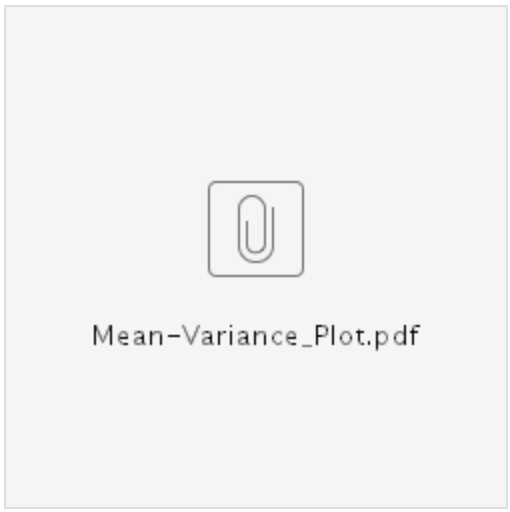
Create a folder named "Class_1" in your home directory. Download all Class 1 files below, and place these files into the "Class_1" folder.

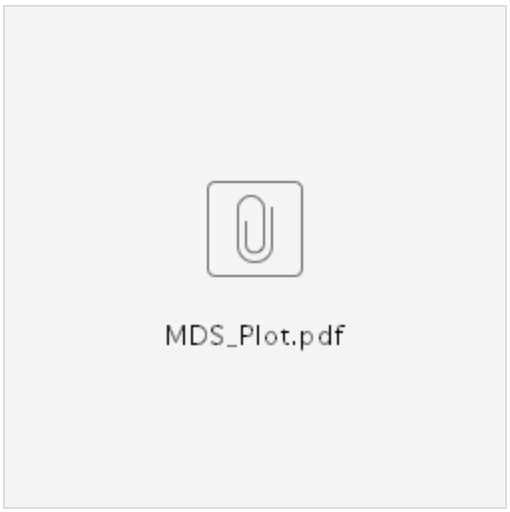


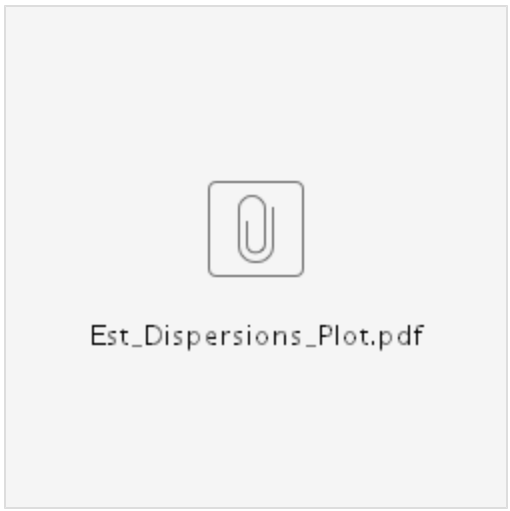
.

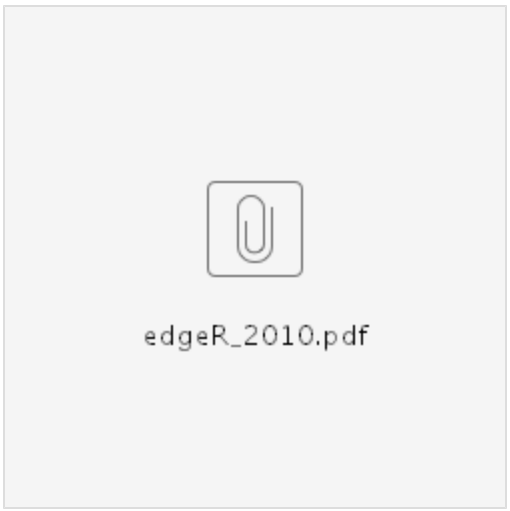








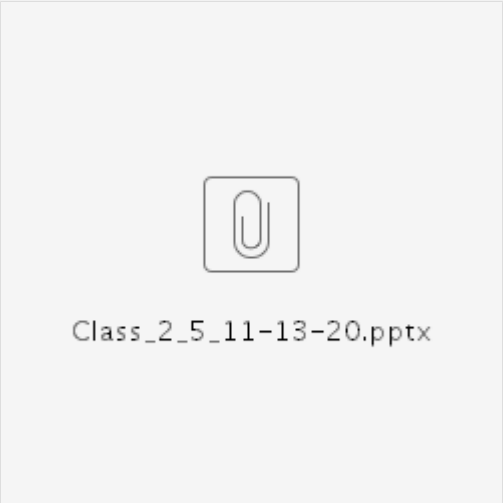


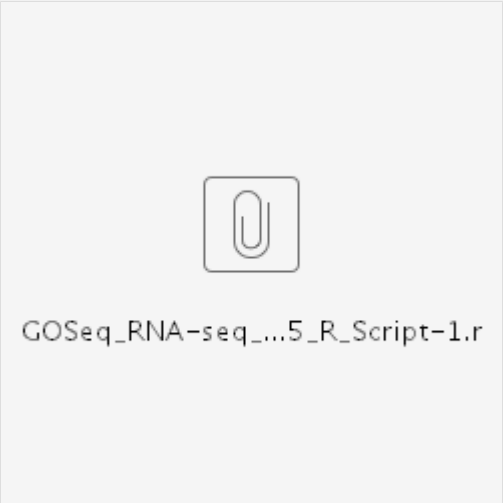


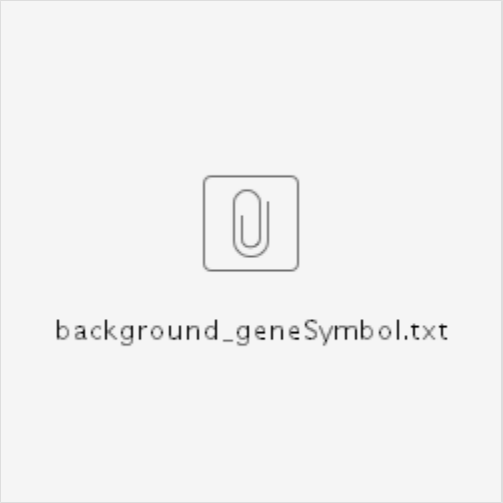
Biostatistics Class 2 Files and Example R Scripts

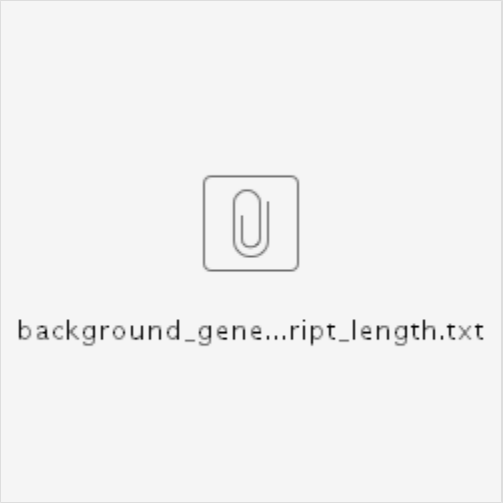
Class 2 covers an introduction to Gene Ontology analysis for RNA-seq and other length biased data. Class 2 includes an in-depth Functional Enrichment analysis of differentially expressed gene lists at varying degrees of false discovery as determined by the edgeR DE analysis in Class 1. Statistical enrichment of gene ontology terms are assessed with the GOseq package by Young et al., 2010 also at the Bioinformatics Division of the Walter and Eliza Hall Institute of Medical Research.

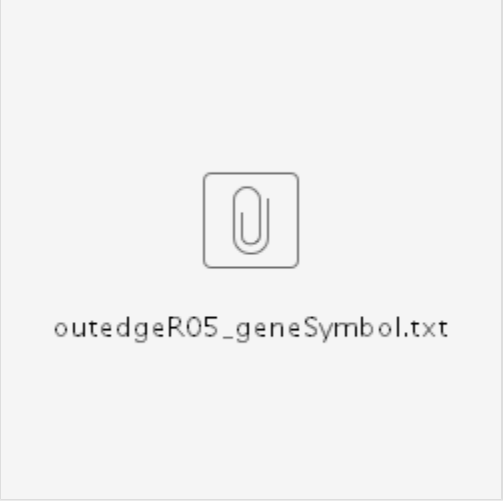
Create a folder named "Class_2" in your home directory. Download all Class 2 files below, and place these files into the "Class_2" folder.

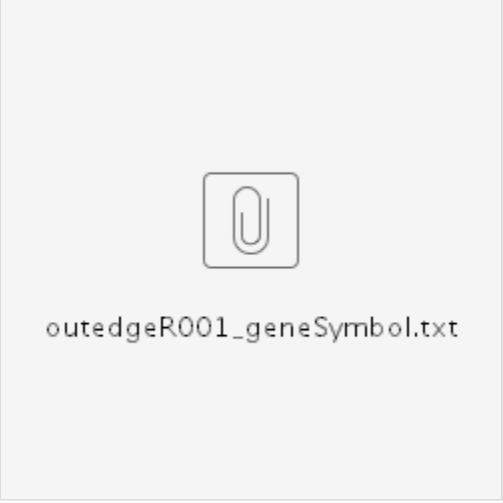


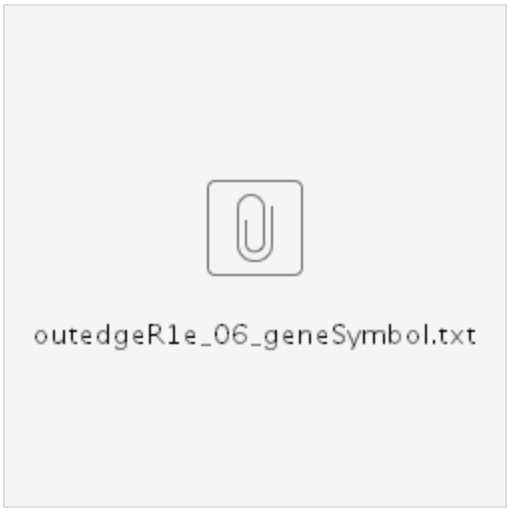


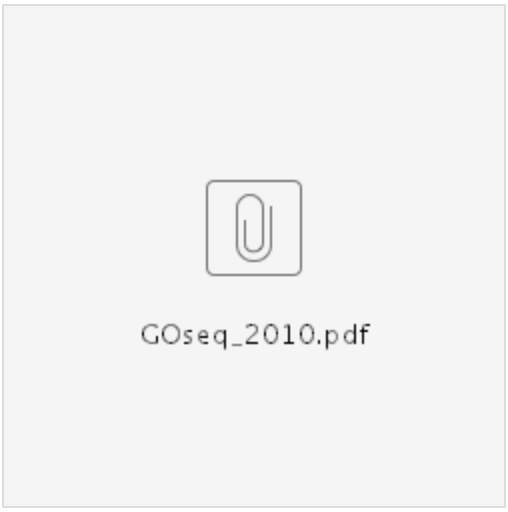


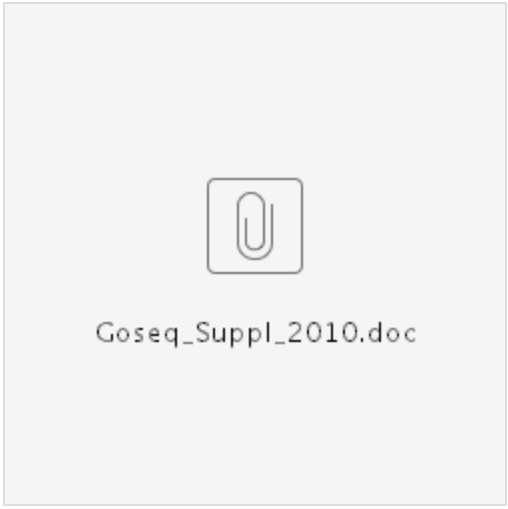












Biostatistics Class 3 Files and Example R Scripts

Class 3 covers Functional Enrichment and Visualization of Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways with the GOSeq package and Pathview, a pathway-based data integration and visualization tool by Weijun Luo and Cory Brouwer at the Department of Bioinformatics and Genomics at UNC Charlotte.

Create a folder named "Class_3" in your home directory. Download all Class 3 files below, and place these files into the "Class_3" folder.

Class_3_6.pptx

GOSeq_RNA-seq_G..._3_6_R_Script.r

background_geneSymbol.txt

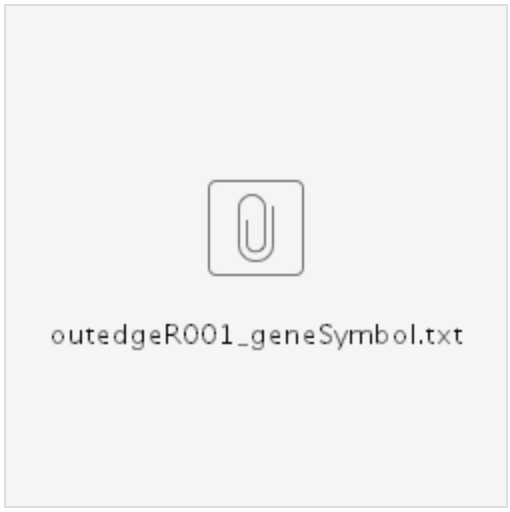
background_gene...ript_length.txt

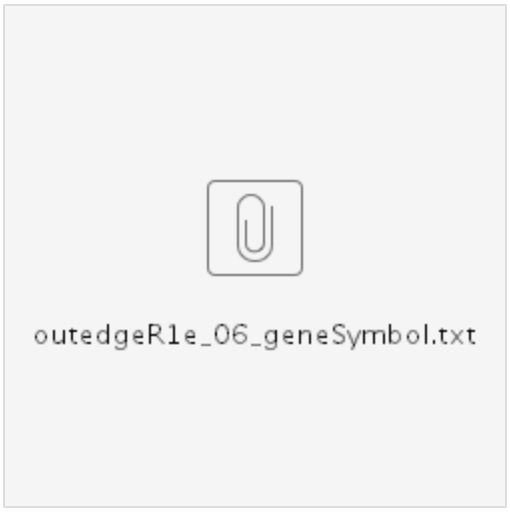
edgeR_tagwise_0..._out_Entrez.txt

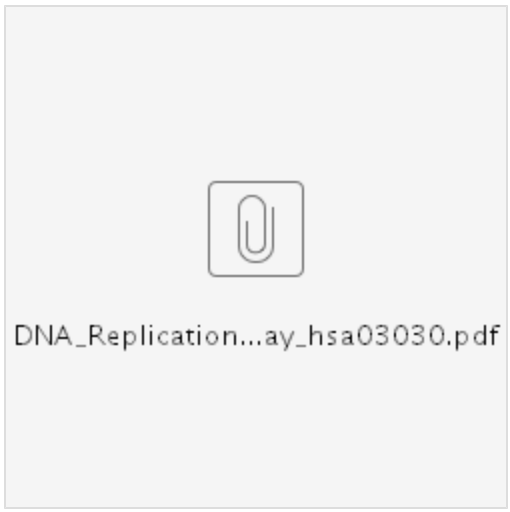
edgeR_tagwise_0..._out_Entrez.txt

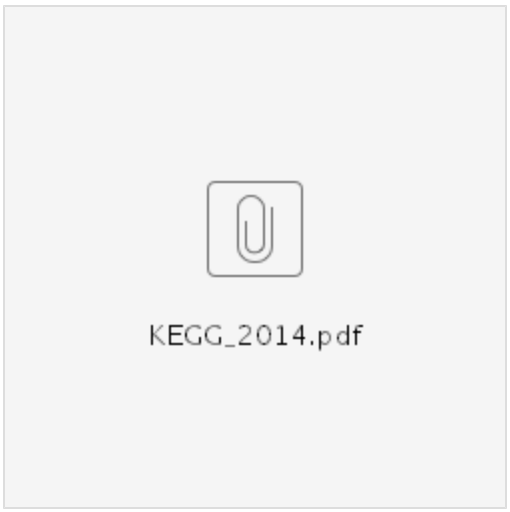
edgeR_tagwise_1..._out_Entrez.txt

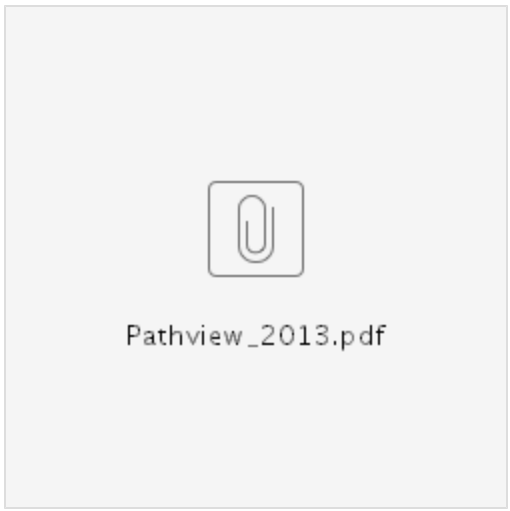
outedgeR05_geneSymbol.txt





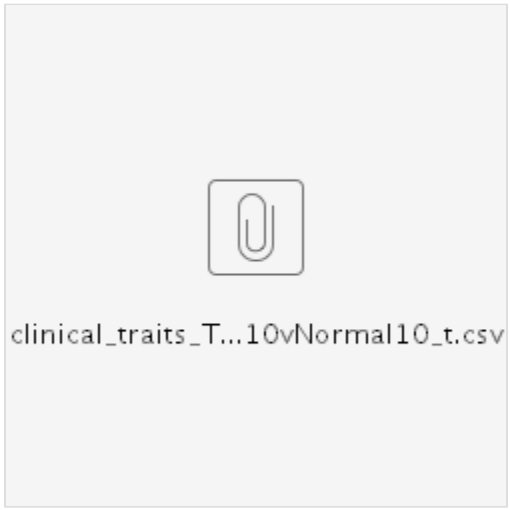


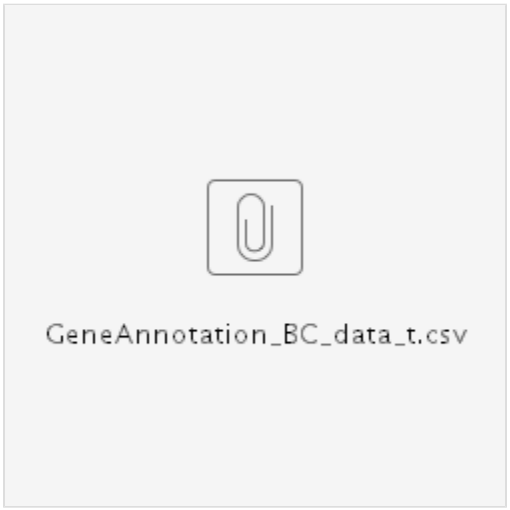


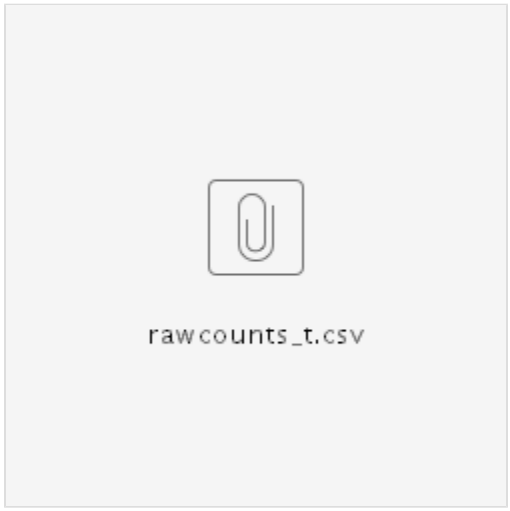


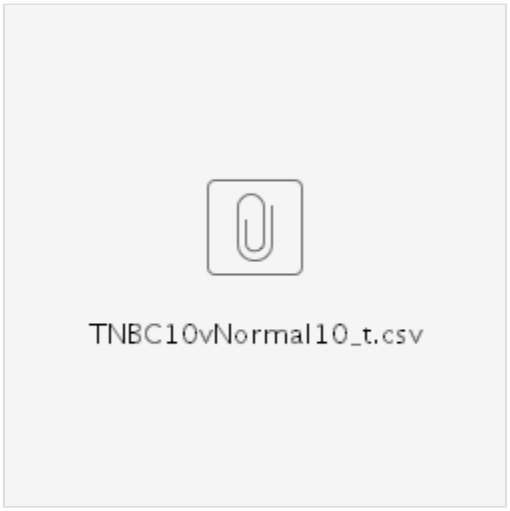
Additional TCGA Breast Cancer Test Dataset

Students are strongly encouraged to apply the biostatistical methods learned in class on the second TCGA Breast Cancer test dataset below.









Session3 PPT:

AI_ASHG_small_10-16-19.pdf

Supplemental Course Information

Gene-specific_D...mation-2007.pdf

Map-Quant_RNA-..._Data_2008.pdf

NB_Exact_Test-C...ersion_2008.pdf

RNA-seq_Length...ounds-2009.pdf

TMM_Normalization-2010.pdf