Harvard School of Public Health

*Department of Biostatistics*

BST 280: Introductory Genomics & Bioinformatics for Health Research

We will analyze gene expression data obtained from of the small intestine of healthy individuals, and whether there are differences between males and females. The data were obtained from The Genotype-Tissue Expression (GTEx) project (<https://gtexportal.org>).

Three files are provided:

1. RNA-Seq gene expression (raw counts) – “gtex\_intestine\_expression.txt”
2. Clinical data, including sex of the individuals – “gtex\_intestine\_clinical.txt”
3. Gene annotation, including Entrez Gene ID and chromosome location – “gtex\_geneAnnotation.txt”

Please download the data, and perform the analysis as outlined below.

1. Genes with very low counts across all libraries provide little evidence for differential expression, and should be filtered out prior to further analysis. Filter genes that have less than 1cpm in half of the samples.
2. Perform a differential expression analysis between males and females. As suggestion, you can use the limma package including the voom method. Report the number of differentially expressed genes under different cutoffs.
3. Determine which GO terms and which KEGG pathways are over-represented in the set of genes highly expressed in males and in the set of genes highly expressed in females under the threshold of adjusted p-value < 0.05. As suggestion, you can use the functions “goana” and “kegga” from the limma package. For this, you need to convert Ensembl IDs to Entrez gene ID. Notice that Entrez gene annotation is not found for every Ensembl ID.