Example BIOS 7659 Projects:

Example 1: Since my lab studies pharmacological response in bladder cancer, I have access to many CEL files of bladder cancer gene expression. I also have access to the same cell lines' GI50 drug response scores for three different drugs commonly used for treating bladder cancer: Gemcitabine, Paclitaxel, and Cisplatin. I want to explore differential gene expression signal in the sensitive versus resistant cell lines, and then check if that signal exists in TCGA bladder cancer patients' genomic profile as well. I plan on normalizing the bladder cancer gene expression cell lines CEL files by using the Affy suite to open and analyze the raw data. Then I will use the expresso package to normalize the data. Next, I will use samr and limma to check differentially expressed genes, and hopefully some genes will overlap between the two methods. Then I will take the TCGA data and separate the patients by those with differential expression in those genes to check the survival trends using a Kaplan Meier plot. Hopefully, I will see that the patients harness the same differential gene expression signature that the cell lines are expressing. I will have three sets of results: one for each drug.

Example 2: I will be working with a mouse RNA-Seq time series data with 13 experimental conditions with 5 biological replicates in each condition; two treatments (air control vs. cigarette smoking mice) at each of six time points (1 day, 7 days, 1 month, 3 months, 6 months, and 9 months), and an additional cessation group at month 9 that had been exposed to 6 months of cigarette smoke and 3 months of air ((2x6) + 1 = 13). Previous analysis of this data has examined differential gene expression at individual time points and across the time series. For the project, I will employ DEXSeq, a package developed by the creators of DESeq, to detect differential exon usage (DEU) and each of individual time points, and in the pairwise comparisons at month 9 (air control vs. cigarette smoke, air control vs. cessation, and cessation vs. cigarette smoke). DEU is a more general concept than alternative splicing, because it also detects alternative transcript start sites and polyadenylation sites. DEXSeq needs aligned reads using a short-read alignment tool (such as GSNAP) that can accommodate reads spanning across introns. Read counts are then typically quantified with the HTSeq package, which can accommodate exon-bins within multi-exon genes. The data were initially prepared with these packages, so only minor adjustment to the package inputs will need to be made. The basic approach of DEXSeq is to calculate the ratio of exon mapped read counts (delineated by coordinate on multi-exon genes) to mapped read counts for the entire gene. The ratios are then compared across conditions to make inference about relative exon usage.

Example 3: For my project I will be conducting an analysis of epigenetic data that studies the biological mechanism of the HPV virus that causes cervical cancer. The study includes four cell types where three have been infected with different types of HPV. Specific interest is in the role of E7, a small protein related to p53 that has various functions where knock down of the gene has been found to stop the progression of cancer. NIKS cells with HPV-16 were compared to NIKS cells with HPV-16 with no E7 along with NIKS cells infected with HPV-18. Differential methylation due to infection of HPV-16 will involve comparison of NIKS with HPV-16 to control cells (NIKS) where differential methylation due to gene E7 will involve comparison of NIKS HPV-16 with NIKS HPV-16 Δ E7 (no E7 gene present). Methylation data from Illumina 450K arrays are in the form of IDAT files in which quality control, normalization as well as differential analysis will be conducted. The primary interest is in both hypo and hyper differentially methylated sites for these two comparisons along with gene lists associated with these CpG sites. As I was provided with the raw image files for the Illumina 450K methylation arrays, the first step will involve quality control and normalization. After these steps, a data set with percent methylation for CpG sites for the four cell types (three replicates each) will be generated. Analysis of differential methylation will be conducted for the single CpG sites where most of the methods discussed in class have involved identification of differentially methylated regions.