Bonus MATLAB bioinformatic section modified from:
Guide to Gray lab afternoon session on analysis of gene expression

First we need to load in the data.
Drag the .tsv file into the MATLAB command window.
This will automatically try to open it with.
You will see something like
>>uiopen('/Users/ms182/Teaching/GrayBootCamp2012__mm9-(Total1,Bulent)(0hr,1hr,6hr)__(EXN,INT)-EXPRESSION.tsv',1)

Unfortunately this won't do exactly what you want but it gives you the proper path to the file. Using copy and paste to help you save all the string between the single quotes into the variable Filepath. This will of course be different but simlar to below and will depend on where on your computer the file is stored.

Filepath='/Users/ms182/Teaching/GrayBootCamp2012__mm9-(Total1,Bulent)(0hr,1hr,6hr)__(EXN,INT)-EXPRESSION.tsv';

Now type:
x=importdata(Filepath);
You have created a structure named x
More on structs later but for now double click on x in the workspace.
It has two fields – data and text data.
Double click each of them and look at them and you should be able to figure out what MATLAB has done.
To manipulate either of these field you would treat it just like any other array but the name would be x.data or x.textdata respective.
For example x.data(:,5) would return the whole five column of the data array in x.
With this knowledge you should be able to make progress on the below exercises.

An experiment was performed where KCl was added to neurons and the cells were collected; 0, 1, and 6 hrs in duplicate. RNAseq was performed on the data. The reads were aligned to the genome and you have a table which lists the number of counts for each gene (each row is a gene) in the exon (EXN) and intron (INT). EXN_num is the number of exons in that gene. EXN_bp and INT_bp are their respective total lengths. Nrds means number of reads. Rep means replicate. We will ignore the density columns.

Exercises

1.

(1) Data quality. How reproducible is the data? Plot and quantify (due this for the exon but write as a function if you can and scripts if not so you can easily redo for introns)

a. Compare all replicate time points

b. Think about what types of plots would be the best way to show off your data.
c. Why is there an offset?
d. How can you correct (normalize) for this offset? Do it.
e. Could this bias your results, in what way?
f. Check out the hint section for ideas of other plots.
g. Calculate the stdev of the replicate measurements as a whole
h. Caluclate the stdev of each replicate measurement and plot this versus average expression level.
i. Try smoothing this data with the smooth function (see help).
j. What do you learn from this?
k. RNAseq is fundamentally a counting method. There is always error due to counting. Assume that the number of counts that you got is correct. Now based on the total number of counts and the frequency of counts in each gene simulate your RNAseq experiment 100 times. Then plot the 95% convidence interval for each gene (based on the average counts).
l. How much of the error is count noise versus other noise?

(2) Converting to gene expression level

a. We have counts but this is not the same as gene expression level. Why not?
b. What do we have to do to convert to expression level?
c. Do it!
d. Now go back through part 1 and see if this affected your data quality. Why/why not?

(3) What genes are neuronal activity-regulated? (RNA-Seq data)

a. First, an analysis of gene expression for any individual timepoint (e.g., no KCl treatment). What are the expression levels of each gene?
    i. Try to produce a plot that shows the expression level of **all** genes. Is gene expression normally distributed? Why or why not? What is the best way to plot the data? How do you deal with genes that have no detectable expression?
    ii. What are the 5 top expressed genes? How many genes are not expressed at all? Are there genes that are detected but not believably expressed?
    iii. For one very highly and one very lowly expressed gene, examine the evidence for its expression in data and in the genome browser (instructions below). For the low expressor, does the evidence support expression of the gene?
b. Now, which genes are activity-regulated?
    i. Produce plots that summarizes the changes of all genes with KCl treatment at one or six hours.
    ii. For the top 10-20 most induced genes, plot their fold induction with confidence intervals. Try making a volcano plot (google it).

iii. How many genes change significantly? For any given gene, how do you know that the difference is not due to random chance?
iv. Do approximately similar numbers of genes go up versus down with activity? Why or why not?
v. For a gene that changes significantly with KCl, observe the raw RNA-Seq data using the UCSC genome browser. What do you see?
vi. Look at either *Homer1* or *Pde10a* in the genome browser (instructions below). What does this example reveal about our analysis?
vii. Based on the above results repeat this section looking at changes in intron expression instead of exon expression. How different does this look?

Genome browser instructions

(1) Point your browser at http://genome.ucsc.edu/.
(2) Click on "Genomes" in the menubar at the top.
(3) Choose the mouse genome, July 2007 assembly. Do you know what an assembly is?
(4) Click "add custom tracks."
(5) Paste in the RNA-Seq KCl experiment data track lines from the file "KCL_genome_browser_tracks.txt". You should see twelve tracks appear, with two replicates of 0, 1, and 6 hours of KCl treatment. There are separate tracks for positive and negative genomic strands (Watson and Crick).
(6) Click "go to genome browser". You will be directed to an arbitrary genomic position. If you enter a gene name in the "gene" box, you can pull up a gene of interest.
(7) You can zoom out with the appropriate buttons and zoom in by dragging your mouse across the upper-most section of the genome image.
(8) Note that your chromosomal position will be indicated in an image of the current chromosome near the top of the page.
(9) How to read the browser:

The twelve tracks are shown on top. A vertical line represents count. The intensity of those plots is the height of the line (log scale!). At the bottom is the different prediction for where exons (thick lines) and introns (thin lines) are for the gene. Beneath that is data on conservation.