**Genetics Bootcamp**

**Guide to Gray lab afternoon session on analysis of gene expression**

**Required files:**

<http://neuroduo.hms.harvard.edu/GeneticsBootCamp/>

RNA-Seq data and exercises contained in:

Processed data:

[GrayBootCamp2012\_\_mm9-(Total1,Bulent)(0hr,1hr,6hr)\_\_(EXN,INT)-EXPRESSION.tsv](http://neuroduo.hms.harvard.edu/GeneticsBootCamp/GrayBootCamp2012__mm9-%28Total1,Bulent%29%280hr,1hr,6hr%29__%28EXN,INT%29-EXPRESSION.tsv) ; includes:

Code:

[Gray\_Lab\_BootCamp2012--FAKEDATA--Gene\_Expression.R](http://neuroduo.hms.harvard.edu/GeneticsBootCamp/Gray_Lab_BootCamp2012--FAKEDATA--Gene_Expression.R)

Fluidigm microfluidic qPCR data for several KCl experiments:

Raw Data:

Fluidigm\_Chip1\_01\_01\_2011.tsv

Fluidigm\_Chip2\_03\_07\_2011.tsv

Fluidigm\_Chip3\_04\_12\_2011.tsv

Code:

Fluidigm\_Bootcamp\_Required\_Functions.R (load these required functions)

Fluidigm\_Bootcamp.R (this is your script to follow).

NanoString digital hybridization data for two KCl experiments:

Raw Data:

NanoString\_KCl.tsv

Code:

nanoString.R (this is your script to follow)

**Exercises**

1. **What genes are neuronal activity-regulated? (RNA-Seq data)** 
   1. First, an analysis of gene expression for any individual timepoint (e.g., no KCl treatment). What are the expression levels of each gene?
      1. 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?
      2. How reproducible is the data? Plot and quantify.
      3. For each gene, generate a mean and 95% confidence interval. What kinds of variability does this error reflect?
      4. 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?
      5. 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?
   2. Now, which genes are activity-regulated?
      1. Produce plots that summarizes the changes of all genes with KCl treatment at one or six hours.
      2. For the top 10-20 most induced genes, plot their fold induction with confidence intervals.
      3. How many genes change significantly? For any given gene, how do you know that the difference is not due to random chance?
      4. Do approximately similar numbers of genes go up versus down with activity? Why or why not?
      5. For a gene that changes significantly with KCl, observe the raw RNA-Seq data using the UCSC genome browser. What do you see?
      6. Look at either *Homer1* or *Pde10a* in the genome browser (instructions below). What does this example reveal about our analysis?
2. **Dissect the diversity of the activity-regulated gene program. What can we learn about this program just by looking at it more carefully (NanoString and Fluidigm data).** 
   1. Start with the NanoString data, which contains in columns 4-6 a repeat of the RNA-Seq KCl experiment (0, 1, 3, 6 hours of sustained KCl). First plot the induction of a single gene that you found to be induced in the RNA-Seq experiment. What kind of normalization, if any, is appropriate here?
   2. Now for the same experiment, produce and save a set of plots for all the genes tested using NanoString. What do you conclude?
   3. Now plot induction using the Fluidigm data. Are the two platforms in agreement? How many induction patterns do you see? For convenience, pick a gene or two from each category as representative. If you’re looking for more challenge beyond the scope of this session, use the function kmeans to answer this question using “unbiased’ clustering.
   4. Why is the induction of some genes slow? Are the slow genes simply less induced per unit time during which KCl is present? Test this with a short pulse of KCl (Experiment 2 in the NanoString data).
   5. Do some of the slow genes require an initial wave of transcription factors to be transcribed and translated? Answer this question using the “CHX/ANS” experiment in which translation was blocked with CHX or ANS during the KCl treatment.

**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.