Linking peaks to genes in Galaxy (modified for classroom)

You have found MACS peaks, now you'd like to see if and how they might overlap with annotated genes. Note that you could select any number of other gene annotation sets, ESTs or (importantly) just about any other track in UCSC (e.g. histone modifications, repetitive elements, microRNAs, or any other). But we will start with Refgenes.

1. Upload mm9 mouse Refseq gene coordinates from UCSC
    a. Get_Data ->UCSC_Main Table Browser
    b. When the browser window opens, select Mouse, July 2007 (Mm9) as your genome (since that is what we mapped the bowtie reads to)
    c. Group: "Genes and Gene Predictions", track: "Refseq genes", table "Refgenes", region: "genome"; output format "Selected fields from primary and related tables"; Be sure "Send output to Galaxy" is clicked. Other fields are left as set by default. When the next window opens:
        Select the following boxes:
            Name (this is transcript Identifier)
            Chromosome
            Strand
            Tx start
            Tx end
            Name2 (this is the common gene name)
    d. Select" get output" to send the file to your Galaxy account.
    e. Take a look at the file by clicking on the eye and note down which columns contain which features (e.g. chromosome, start, end, strand and identifiers (column 1 is the unique id for each transcript).
2. Now, for what we want to do we need to alter the file type and attributes of this file; Galaxy may have saved it in the tabular format but you need .bed. You will click on the pencil in the right side panel to do this.
    a. First select the "Datatype" tab and select "Bed" as the format. Select "save".
    b. In the pulldown menus that appear, tell Galaxy which columns contain the various features. Deselect "Score column" since this file has none.
    c. Click save.
3. Now, to intersect the two files:
    a. Under "Operate on Genomic Intervals", select "Join the intervals of two datasets side by side"
    b. For the first file, select your "Macs on data x and y (where x was the CTCF file and y was the input control) (peaks:bed).
    c. For the second file, select the mouse Refgene file
    d. Leave "with minimal overlap" at 1 bp
    e. Return: select "All records of the first dataset".
    f. Click "execute".
    g. Click the eye and check out the intersecting file. This should be a list of all your peaks, and for those with an overlap to Refgenes, a list of the Refgenes. If the peak overlapped more than one Refgene/transcript it will be listed more than once.

4. There are a lot of ways you can modify this pipeline. For example, some of the peaks do not overlap genes, but you'd like to know what genes they might be near (e.g. the might be just upstream or downstream).

   One way to do this is to use the "text manipulation tool" to create a modified version of the peaks: bed file. First look at this file, and record which columns have which information.

   i. Select "compute an expression on every row"
      1. Add expression: c2-10000 (will extend the most telomeric coordinates of your peaks by 10K)
      2. Make sure your "peaks:bed" file is selected; select "round result" as "yes"
      3. Press "execute". Now, look at the new file and make sure it worked.
   ii. Repeat the process ON THE NEW FILE, *adding* 10K to the downstream coordinate as a second new column. This new file should have two columns that together expand your peak coordinates by 20 kb.
   iii. Edit attributes of the new file so that the new columns are designated as the "start" and "end" coordinates. Be sure the other columns are properly identified (note there is no strand column here, but there is a score).
   iv. Select "save"

5. Repeat the Refgene intersection with your new file.
   a. How does it differ from the first file?
6. Another similar approach would be to identify the closest gene(s) upstream and/or downstream of your peaks, using "Fetch closest non-overlapping feature". For this you should use the original "peaks:bed" file.
   a. Try it, selecting "both upstream and downstream", check the output.
   b. Explain, in a few sentences, how this approach and the one in 4 above could (and should) give you different types of results

Note that you can use this same approach to find features of any kind that overlap with, or closely flank, your peaks. For example instead of 'Refgenes" (or any other gene annotation file) you could have downloaded and used a table with Repetitive elements ("Repeatmasker"); ENCODE features of many different types (histone modifications; TF binding regions, etc etc etc).