Before we start:

1. Log into tak (step 0 on the exercises)
2. Go to your lab space and create a folder for the class (see separate hand out)
3. Connect to your lab space through the wihtdata network and open the folder you created

Questions you may have after a ChIP-seq, DHS or similar type of experiment

Given a set of genomic regions of interest, how can one identify
- binding preferences (promoters, exons, introns, intergenic regions)
- genes that may be regulated by proteins binding to those regions
- sequences for motif analysis
- CpG island content
- SNPs overlapping with the regions (SNPs could affect regulation)

From genomic regions to biology

BaRC Hot Topics – March 2014
http://jura.wi.mit.edu/bio/education/hot_topics

Tools that will help you answer those questions

- Cis-regulatory Element Annotation System: Ceas tool
  http://liulab.dfc.harvard.edu/CEAS/
  http://liulab.dfc.harvard.edu/CEAS/usermanual.html
- Bedtools
  http://code.google.com/p/bedtools/
  BEDTools: a flexible suite of utilities for comparing genomic features
Data we will use in our hands-on exercises

- Peaks from ENCODE H3K4me3 ChIP-seq in HepG2 cells
  peaks.bed
  - chr1 19921 20036 MACS_peak_1 50.12
  - chr1 20025 20890 MACS_peak_2 568.43
  - chr1 27627 27764 MACS_peak_3 76.25
  - chr1 27976 28955 MACS_peak_4 5289.35

- Wig files generated from MACS from that same experiment
  Hepg2H3k4me3_treat_afterfiting_all.wig
  track type=wiggle_0 name="Hepg2H3k4me3_treat_all" description="Extended tag pileup from MACS version 1.4.2 20120305 for every 25 bp"
  variableStep chrom=chr1 span=25
  10026 2
  10076 1

- Bed file with ENSEMBL genes
  GRCh37.p13.HumanENSEMBLgenes.bed
  - chr1 50927700 50927800 ENSG00000271782_RP5C850O15.4
  - chr1 103817769 103828355 ENSG00000232753_RP11C347K2.1
  - chr1 50927741 50927822 ENSG00000237576_RPS50D513.3
  - chr1 50965430 50965529 ENSG000000202140_Y_RNA

- Bed file with CpG islands in the human genome (downloaded from UCSC)
  cpgIslandExt.bed
  - chr1 28735 29833 Cpg_116
  - chr1 315124 315136 Cpg_120
  - chr1 327790 328229 Cpg_29
  - chr1 437511 448316 Cpg_34

Exercise 1: CEAS tool

- Input files
  - Bed file with regions (ChIP-seq peaks)
  - Wig files: ChIP-seq signal
  - Genome annotation files that can be downloaded with the tool
    Location of the genome files: /nfs/BaRC_datasets/CEAS/genome

- Run exercise 1 command:
  bsub --ceas -b peaks.bed -w Hepg2H3k4me3_treat_afterfiting_all.wig -g /nfs/BaRC_datasets/CEAS genome/hg19.refGene

Output from ceas tool

- Sample figures from the pdf file
Exercise 2: Find genes that may be active on this sample (have H3K4me3 peaks close to them)

- Take all genes and add 3Kb up and down with slopBed
  ```bash
slopBed -b 3000 -i GRCh37.p13.HumanENSEMBLgenes.bed -g /nfs/ genomes/human_gp_feb_09_no_random/anno/chromInfo.txt > HumanGenesPlusMinus3kb.bed
  ```
- Intersect the slopped genes with peaks and get the list of unique genes overlapping
  ```bash
  intersectBed -wa -a HumanGenesPlusMinus3kb.bed -b peaks.bed | awk '{print $4}' | sort -u > GenesAt3KborLessFromPeaks.txt
  ```

Exercise 3: A different way of finding genes next to peaks

For each region find the closest gene and filter based on the distance to the gene

```bash
closestBed -d -a peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10 -c 6,7,8, -o distinct,distinct,distinct | awk '{BEGIN {OFS="\t"}; if ($2<3000) {print $3,$4,$5,$1,$2}}' > closestGeneAt3KborLess.bed
closestBed -d print the distance to the feature in -b
groupBy
-g columns to group on
-c columns to summarize
-o operation to use to summarize
```
Exercise 3: Step by step
For each region find the closest gene and filter based on the distance to the gene.

closestBed -d -a peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | head

Go to Exercise 4.

Exercise 4: Get the sequence of the regions of interest

bsub "fastaFromBed -fi /nfs/genomes/human_gp_feb_09_no_random/fasta_whole_genome/hg19.fa -bed peaks.bed -fo peaksSeq.fa -name "

fastaFromBed
-fi Input FASTA file
-bed BED/GFF/VCF file of ranges to extract from --fi
-fo Output file (can be FASTA or TAB-delimited) --name Use the name field for the FASTA header

If the regions have a strand information you can use -s:
it will extract the sequence in the orientation defined in the strand column

Exercise 5: Count the number of CpG islands overlapping with the regions

coverageBed -a cpgIslandExt.bed -b peaks.bed > cpgIslandsCountsInPeaks.bed

cpgIslandsCountsInPeaks.bed

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>end</th>
<th>peakName</th>
<th>-10Logpval</th>
<th># of a</th>
<th>length of a</th>
<th>length of b</th>
<th>% bases of b covered by a</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1 9922640 9994747</td>
<td>MACS_peak_1443</td>
<td>1152.98</td>
<td>1</td>
<td>130</td>
<td>831</td>
<td>0.156438</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr1 135141 135945</td>
<td>MACS_peak_1728</td>
<td>801.18</td>
<td>1</td>
<td>69</td>
<td>764</td>
<td>0.090141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr1 1796240 1798079</td>
<td>MACS_peak_992</td>
<td>715.58</td>
<td>1</td>
<td>23</td>
<td>684</td>
<td>0.0336257</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*header is not part of the "cpgIslandsCountsInPeaks.bed" file, it is added to the table to clarify what each column is
Other resources

• Previous Hot Topics
  – http://jura.wi.mit.edu/bio/education/hot_topics/

• Hot Topics scheduled for April and May
  – Ngsplots
  – python