1. Exercises: Using the UCSC Genome Browser

In this section, you will learn how to set up a Genome Browser session which shows only the tracks that are of interest to you. For this exercise, we will ask you to display:

- *Genes*
- *Single nucleotide polymorphisms (normal variances among individuals, known as SNPs)*
- *CpG islands (parts of the genome involved in regulating transcription)*
- *Conservation (a measure of the degree of similarity of a genomic region between species)*

You will then search for a dataset from the ENCODE database which contains data from an experiment measuring the 'H3K4me3 histone modification' (a marker for 'promoters' which are regulatory regions at the start of genes) in the cell-line known as 'NHEK'. Once you have put together your UCSC session using publically available data, we will then move on to the next exercises – using Galaxy and uploading custom datasets.

1. Load the following UCSC track into your browser – user: rpoulos & session name: workshop.

2. Add the "UCSC Genes" and "RefSeq Genes" tracks in full. What differences are there between the two tracks? After answering this question, you can hide the "RefSeq Genes" track.

3. Add the "Common SNPs(147)" track and ensure that all SNPs are shown in a single line, rather than descending across the screen.

4. Is there a CpG island present? If so, where is it located?

5. What are the most conserved parts of the region displayed on your browser?

6. Click on "track search" and search for "NHEK H3K4me3". What tissue do NHEK cells come from? Add the tracks from the ENCODE database (Broad Institute) to the browser.

7. Configure the "NHEK H3K4m3" track. Change the track height to 100 pixels and set the vertical viewing range maximum to 150.

8. To check your results, load the following UCSC track in a separate window and compare your browser output with that shown in the following – user: rpoulos & session name: example

2. Exercises: Using Galaxy to perform simple genome arithmetic

In this section, you will use Galaxy to find regions of the genome which fulfil certain criteria. You will be working with a list of mutations from a whole-genome sequenced melanoma cell-line known as "COLO829". We have provided you with a file listing the COLO829 mutations. We have also given you a file listing regions containing the NHEK H3K4me3 histone modifications that we looked at in the previous exercise. You will need to upload these files, and then import additional files from UCSC. We will ask you to import two files from UCSC - one which contains the genomic coordinates of all simple repeats in the genome (simple repeats are runs of bases which are commonly repeated) and one which contains all CpG islands in the genome. For this analysis, we want to find high confidence mutations which lie within regulatory regions at the start of genes. First, we will ask you to remove any mutations from the COLO829 cell-line which lie within simple repeats, as mutations in these regions are commonly false positives. Using your new list of mutations, we will then ask you to find which mutations overlap both CpG islands and peaks in the NHEK H3K4me3 dataset. This will give you a list of high confidence COLO829 mutations which lie within regulatory regions at the start of genes. We will then ask you to download this list and use it for the next exercise.

1. Upload datasets onto Galaxy by using the tools listed under "Get Data" in the left panel.
   - Import the following dataset from the UCSC Main table browser:
     - "CpG Islands", search using the "All Tracks" option
   - Upload the files "COLO829 mutations.bed", "SimpleRepeats.bed" and "NHEK_H3K4me3_peaks.txt" which were provided for you. Set the Genome for these datasets as "Human Feb. 2009 (GRCh37/hg19) (hg19)".

2. View the COLO829 mutations dataset. How many columns are in the file? Which columns give the mutation coordinates?

3. Perform a line count on each of the files.

4. How many COLO829 mutations do not overlap simple repeats?
   Hint: Use one of the tools listed under the "BEDTools" option.

5. Using the list of COLO829 mutations that do not overlap simple repeats, determine how many mutations are within CpG Islands and H3K4me3 peaks in NHEK cells.
   Hint: Use the "Intersect" tool within the "Operate on Genomic Intervals" list.

6. Download the file containing this final list of COLO829 mutations.

3. Exercise: Integrating UCSC & Galaxy

In this section, you will use the UCSC Genome Browser session that you created earlier, and integrate the dataset that you obtained from Galaxy. First, you will upload the list of COLO829 mutations that fulfilled your specified criteria, and view one of the mutations from the list on the UCSC Genome Browser. Then you will upload your list of all COLO829 mutations and view a specific region of the genome, and provide some information about the mutations in that field of interest.

1. Return to your UCSC session. (If you have closed this session, you can reload the session using user: rpoulos & session name: example).

2. Select "Add Custom Tracks" and upload the file that you downloaded from Galaxy at question 6 in the previous exercise. Click "Submit". In the new page which loads, select "User Track" under the "Name" option. In the new page which loads, under "Edit configuration", specify a new track name. Also, add in the option "color='100,0,100'" (this will make your mutations appear in the browser in a specific colour – in this case purple, rather than black). Return to the genome browser.

3. View the region which lies between the following coordinates chr8:125,540,555-125,560,554. Can you see a mutation in this region? Describe its location.

4. Zoom into the position of the mutation. What nucleotide is mutated (A,C,G or T)? How conserved is the mutated base across species?

5. Upload the original "COLO829_mutations.bed" file to your session. Specify a new track name and change the colour to green (0,100,100). Set the track in the genome browser to view under the "Pack" option.

6. Zoom to the 10,000 bp region at chr10:28,027,012-28,037,011. How many of the filtered COLO829 mutations (from the first file) are in this region? How many total COLO829 mutations are present in this region? Of the total COLO829 mutations, what types of mutations are they? (ie give the wild-type and mutant bases). In what gene do they lie? What is the name of the part of the gene in which they lie?

4. Exercise: Unguided exercise

In this section, you will need to apply your knowledge from the lecture and exercises to solve a problem using UCSC and Galaxy. Answers and step-by-step instructions to solving this problem have been provided in the manual.

Task:

Determine how many nucleotides are listed as both:

- Positions for variants from the ClinVar database
- Being within DNase I hypersensitivity clusters with scores greater than 500

(There are many tracks which identify different measures of DNase I hypersensitivity in different cell-types. For this exercise, use the “DNase clusters (v3)” track listed under “ENCODE Regulation”).