1. Answers: Using the UCSC Genome Browser

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

   Click "My Data", then "Sessions". Scroll to the "Restore Settings" heading and enter the user and session information supplied above.

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.

   You will see two genes. One gene (TATDN1) starts in the centre of the region and runs from right to left (3' to 5'). The other gene (NDUFB9) starts in a similar location and runs from left to right (5' to 3'). RefSeq gives many different transcripts for these genes (with differing exons & introns), but UCSC only gives one transcript.

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

   Set the track to the "dense" mode.

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

   Show the track "CpG Islands…" in your browser. There is a CpG island present which bridges the start of the TATDN1 and NDUFB9 genes. Click on the CpG Island to give more information – it's coordinates are chr8:125,551,212-125,551,521.

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

   Show the "Conservation" track in full on your browser. The most conserved regions coincide with the exons of the NDUFB9 gene.

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.

   Expand the arrow and click on "NHEK". You will see the "Tissue" type listed as "skin". You will be adding 2 tracks, both named "NHEK H3K4m3" 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.

   To do this, right click on the track and select "Configure NHEK H3K4m3". Change the relevant settings in the dialogue box which will open up.

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. Answers: Using Galaxy to perform simple genome arithmetic

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:
     - o “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?

   To view the file, click on the eye next to the COLO829 mutations dataset in the panel on the right. There are 4 columns in the file. The first 3 columns give the mutation coordinates.

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

   In the left panel, click on “Text Manipulation” and then the “Line/Word/Character count of a dataset” tool. The first column of the output file will provide the line count.
   COLO829 mutations – 33,340 lines
   Simple repeats – 962,714 lines
   CpG Islands – 28,691 lines
   NHEK H3K4me3 – 51,310 lines

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

   Use SubtractBed, with COLO829 mutations as file A and UCSC Simple Repeats as file B. There are 32,584 COLO829 mutations which do not overlap simple repeats.

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.

   You will need to do this intersection in 2 steps. If you intersect with CpG Islands first, you will find 119 mutations within CpG islands. If you intersect with NHEK H3K4me3 peaks first, you will find 733 mutations that fall within these peaks. The final count of COLO829 mutations which are not within simple repeats but are within CpG islands and within NHEK H3K4me3 peaks is 73 mutations.

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

   Click on the file in the right panel and select the image of the floppy disk to download the file. Save this file to your hard drive.

<u>3. Answers:</u> Integrating UCSC & Galaxy

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 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.

   There is one mutation which lies within the CpG island and the NHEK H3K4m3 peak. It overlaps the beginning of the TATDN1 and NDUFB9 genes.

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?

   The mutated nucleotide is a C. The base is conserved across all species shown from Human through to Elephant. This is because all genomes show a C in that position.

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?

   There is one mutation which fulfils the filtered criteria. There are two COLO829 mutations in total. These mutations are of the type G>A and C>G. (Note that the mutation type is shown in the UCSC Genome Browser since the data of mutation type was listed in the fourth column of the uploaded file). Both mutations are within introns of the MKX gene.

   To check your result for this question, you can load the session user: rpoulos & session name: example2.

4. Answers: Unguided exercise

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").

Below are some step-by-step instructions for completing the given task. There are a number of different tools/pipelines which you could use to determine the correct answer. Do not worry if you haven't followed these steps exactly, so long as you end up producing the same final result.

1. *Import the following tracks from UCSC: "ClinVar Variants" and "DNase Clusters". Find them by searching through the "All Tracks" option.*

2. *In UCSC, select "Describe table schema" to work out which column in the DNase clusters file contains the score (you will find that it is column 5). Use the "Filter" tool in Galaxy to filter the DNase clusters file by column 5 for scores ">500", so that the output file contains only these regions. The output file should contain 354,778 variants.*

3. *Use the tool listed under "Operate on Genomic Intervals": "Intersect the intervals of two datasets" to obtain a list of ClinVar Variants that fall into DNase clusters with scores >500. The output file will contain 33,540 variants.*

4. *Your output file from Step 3 contains multiple variants on the same base. In addition, a variant may be listed multiple times if it falls into more than one DNase cluster. To work out the number of unique bases, you will need to perform a "Count" (listed under "Statistics" tools) on this file based on occurrences of values in columns 1, 2 and 3 (ie enter columns 1,2 and 3 under the heading "Count occurrences of values in column(s)"). The final count will be 32,364 bases.*

In conclusion, there are 32,364 nucleotides which are listed as positions for variants in the ClinVar database which also fall into DNase I hypersensitivity clusters with scores greater than 500.