Some of the following is based on a Galaxy tutorial from OpenHelix (www.openhelix.com—galaxy), but it is heavily adapted and extended for The Genomic HyperBrowser

Descriptive statistics

Basic use

- Access the **The Genomic HyperBrowser site** (hyperbrowser.uio.no)
- Choose **User** in the top right corner and choose **Register**. Type e-mail and a password. This will make sure that your data and results will be available for later use on a different computer
- Check that you are logged in by holding the mouse over **User** in the top right corner
- Choose **The Genomic HyperBrowser->Perform analysis**
- Choose **First Track -> Genes and gene subsets -> Genes**
- Choose **Ensembl**
- Choose **Second Track -> -- No Track (single track analysis) --**
- Choose **Descriptive statistics -> Proportional coverage**
- Choose **Region and scale -> Chromosomes**, and leave ***** in the **Which** text box. This instructs the HyperBrowser to run the analysis over all chromosomes. The default option is to only do the analysis on chromosome arms (excluding centromeres)
- Click **Start Analysis**
- Click the **Eye icon** in the new **Perform analysis** history element
- Click on the different **output links** and try to understand what they show. Notice that the **Plot: values per bin** figure only has a straight line per chromosome. This is because of the **scale** of the analysis is only set to whole chromosomes.
- Redo the previous analysis (the **back** button on the browser should work..), but this time, choose **Region and scale -> Custom specification**. Leave ***** in the **Region** box, but put **5m** in the **Bin size** box. This instructs the HyperBrowser to do the analysis in bins of size 5 million base pairs over the complete genome.
- Check the **Plot: values per bin**, and other outputs and notice how they differ.

  Now, with some clicking around, you should be able to answer the following questions:

    - How much of the genome do **Refseq genes** cover?
    - Find the basepair overlap (coverage) between **Ensembl** and **Hinxton Coverage** (notice that this is a two-track analysis). Do the same with **Vega** and **Hinxton Coverage.** What can you conclude?
- Choose **Public tools -> Extract Track -> Genes... -> CCDS**. Choose **Chromosomes -> chr21**. Choose **Segments, any overlaps clustered**. Click **Extract track**.
- The dataset should appear as an item in your **history** on the right side
- **Click the title hyperlink** to examine the summary data
- **Click the eyeball icon** to view the data in the center panel area
- **Click the pencil icon** to be able to prepare the data set for subsequent analyses
- Change the name to: **CCDS genes clustered (chr21)**. You should name important history elements to more easily find them later.
- Click **Save**
- Now, can you find a way to count the number of **Sequence -> Repeating elements** in the genes defined by the CCDS gene track?

Importing custom datasets

- **Click on the "Get Data" link under "Galaxy Tools" in the left Tools area** to expand that topic and see the available data sources. **Click UCSC Main**
- On the right frame the UCSC Table Browser interface will be displayed. Make these changes to the UCSC interface:

  Assembly: Mar. 2006 (NCBI36/hg18)

  In group: **choose Variation and Repeats**, and choose **SNP(131)** as table *(other choices will be set automatically for you)*

  In region: move radio button to **position**, and clear out the default location. **Enter BRCA1** in the text box. Click **lookup**. From the results list **choose the first BRCA1 item** at the top by clicking on it. Its position will display back in the position box

  Leave any other menu settings as default at this time
- Move to the output format area. Leave the menu set to **"BED - browser extensible data**". **Check the box for "Send output to Galaxy"**
- Open the HyperBrowser and choose **-- From history (bed, wig, customtrack) --**. Find the distribution of the length of the segments (do not cluster overlapping segments). What does this say about the contents of the SNP database?

Scatter plot

- In this exercise, you should try to find out what to do more by yourself. Please ask questions if you are stuck.
- The goal of this exercise is to create a scatter plot of the count of **ATG** in the **sequence** on one axis and the number of **genes** on the other. Use any definition. Use **cytobands** as bins. Each bin will define a point in the scatter plot, where the coordinates are the counts of the tracks.
- Are the results surprising?

Hypothesis testing

Virus integration sites

- Choose **Options -> Create New**. This creates a new history for this exercise. You can name the history by clicking the pencil. **Options -> List your histories - Stored by you** gives you the list of all histories. Histories can be shared with other users.
- Choose the **HIV virus integration dataset** from Derse et. al. (under **Phenotypes...**) as track 1
- Choose the **Conserved in Mouse** dataset as track 2
- Choose **Hypothesis test -> Located inside**.
- Choose Treat 'HIV (Virus integration, Derse et al. (2007))' as: **The middle point of every segment...** Note that the viral integration sites are stored as segments, but that we here treat them as their midpoints.
- Choose **more** as the **alternative hypothesis**. This instructs the HyperBrowser to use a right-tailed test.
- Choose **Null hypothesis -> Preserve segments (T2) and number of points (T1), randomize point positions**. This is a analytical test. The other null hypotheses have solutions that make use of Monte Carlo simulations, which runs too long for practical use in these exercises. In this case, making the viral insertion point random, while keeping the rest of the data fixed seems like a natural assumption.
- Use all **chromosome arms** as bins.
- Click on the **information icon** just above the "Start analysis" button. Look at the information. Find the **Solution** header and click the link to the **note**. This contains the statistical solutions used for this test.
- Click **Start analysis.**
- Check the results. Note that there is both a **global result** with a p-value, together with **local results** for each chromosome arm. Here the **FDR-corrected p-values** are the ones to check, as we need to use multiple testing correction (We run one test per chromosome arm). A value under **0.1** is deemed significant, as default, meaning that we accept that one out of ten answers are false positives.
- Do the same test with **HPV** virus.
- Repeat the test, but this time use **Refseq exons** as track 2.
- How do you interpret the results? Does it look like theres a difference in the integration preference of the to virus types?