**B&G Practical 2 Genome databases**

There are numerous genome databases available at different sites. In this practical we are going to look at one site and two types of information. The sites is the NCBI which we looked at in the first practical.

The first part of the practical will look at the human genome on the NCBI. This is very complex, and I do not expected you to remember the details! The intention is that you understand a few basic manipulations of the genome, and the variety of data that is present within it. This is an example of a fully annotated genome, most genomes are not nearly as well annotated. Answer the questions stated in the text below. We will stop after you have done question 4.

**The Human Genome Project**

*Part 1a: Basic information -navigating around the chromosomes and obtaining specific tracks/databases within a particular chromosome*

Sequence information on the human genome is held in several centres, one of the main ones being NCBI. Hence go to:

http://www.ncbi.nlm.nih.gov/genome/guide/human/

Or you could just type in ‘ncbi genomes’ into google and click on the ‘human genome’. There are a lot of ways of accessing the same page.

Click on the Y chromosome in the chromosomes box. A rather complicated display appears! Spend some time investigating this page.

To start: Try zooming in and out the Y chromosome using the – and + sliding bar. Moving the bar to the left shows the whole chromosome whereas to the right zooms so far in that you can see individual bases. (All N’s=unknown bases, why? Surely the human genome has been fully sequenced years ago?) Use the arrow keys on the left of the sliding bar to move around up and down the chromosome. You could try clicking on the tools section also-it is on the right hand side of the ‘slider’

You can also navigate by clicking on the chromosome image towards the top. This image is based on how chromosomes appear in cytogenetics (not much cytogenetics is done these days, but it used to be very popular-spotting chromosomal breakpoints in people with leukaemia for example. The ‘cytogenetics’ display is called a ‘chromosome ideogram’ with some sections showing up as dark bands and some as light bands under a light microscope (google it if you like). Click on the bit you want to have a look at

You can also navigate using the ‘search assembly’ box at the top left. e.g. type in chrY:22M-23M to see that 1 million bp region of the Y chromosome (see the examples button to explore further search criteria).

Question 1. Zoom out so that the whole Y chromosome is displayed again. How many genes are there on the Y chromosome (see the ! Exon navigator) towards the top-it says that there are too many genes to display properly? How many are there?

Question 2. What does lilac area starting at q12 represent? Hint look up euchromatin and heterochromatin in google and room right in to a region to the right of q12

Question 3. How many tracks (libraries) are available to view on the Y chromosome (scroll all the way to the bottom).

Answer there are 825 tracks available. This means that there are 825 different bioinformatic databases annotated with respect to the Y chromosome and other human chromosomes!! Annotation mean that certain sections/regions/bases are associated with one or more of the 825 databases. 11 of the 825 tracks are shown in the default view. Thus 824! more are not displayed in the default view. YOU ARE NOT EXPECTED TO BECOME EXPERTS IN THIS! I am merely drawing your attention to the huge amount of information that is contained in a heavily annotated genome.

Click on the ‘gear’ symbol at the bottom right of the screen next to the 885 number to explore some of them. Again, I don’t expect you to know what each one shows, merely gain an impression of the breadth)! N.B: the human reference genome sequence you are looking at is a composite genome-it is not based on any one individual, but rather chunks of DNA spliced together from lots of different human populations.

We are now going to have a look at just a few of these tracks. Click on the ‘gear cog’ and a new display appears. This allows you to simplify the data on the screen to make it more manageable. If you get lost or make a mistake just hit the ‘Reset tracks’ button at the bottom of the ‘gear cog’ screen and start again.

To start off get rid of all the tracks (by removing the ‘tick symbols’) except sequence and hit ‘configure’ at the bottom of the screen. Much simpler now. Have a play around for 10mins or so by adding and then deleting tracks at random using the menu on the right-hand side. Please bear in mind that there will be some bugs in the viewer I suspect, leading to ‘errors in track loading etc’

Configure the tracks to just show ‘sequence’ ‘tiling path’ and ‘scaffolds’ (you can use the search ‘keyword’ function if you want to find ‘tiling path’ and ‘scaffold’ but they should be at the top)

I want you to look at these so you understand how the genome has been assembled historically. i.e these tracks indicate how the human genome (Y chromosome in this instance) was assembled from smaller sequenced DNA fragments.

Question 4. How is the tiling path (components) and scaffolds related to each other and how do these relate to the accession number of a piece of DNA. Hint, hover your mouse over one of the components or scaffolds and then click on the ‘genbank’ record. You might not be able to easily find the answer to this question (but you could google BAC clone to help you-I will explain it along with the other answers

You can explore the different tracks in your own time if you like but it is rather complicated. In fact, I think the genome annotation is now getting so complicated that it is almost unusable unless you have a specific track in mind, but it is good to have an overview of it. This is the purpose of the first part of the practical. In fact, some of the tracks lead to errors and broken links……showing how difficult it is to maintain a genome with complex annotations.

It is easier to find a particular gene using other methods and then come back to this page to find the information on it if we need to, or navigate to others if we don’t.

*Part 1b: Human Gene Information*

We are now going to use another NCBI site to investigate human genes, this is OMIM-one of the most important databases on human genes and disease.

OMIM has a long list of genes. You can browse by chromosome or find a gene or genetic disease via google searches and then type the name of the disease into the search box.

https://www.omim.org/search/advanced/entry

Type in ‘cystic fibrosis’. There are several genes associated with cystic fibrosis, but we will look at the main one, which is the CFTR gene (number 602421-second on the list). Each of these entries can be regarded as the ‘home page’ of a human genetic gene.

Question 5. What is the chromosomal location of the CFTR gene?

Question 6. How does the gene function?

Question 7. Give a couple of examples of known mutations of CFTR causing disease in humans (Hint: scroll down to the Allelic variants section). Have a look at the population distribution if you like. We will return to allelic variation in the third practical.

Again, this page can be thought of as the ‘home page’ for the gene/disease. In addition to all the text there is a lot of information available in the ‘External Links’ section at the top right of the page. You can explore these links find different information about the gene. We will now look at a few examples, but it would be good if you could explore some of the others in your own time.

Question 8. What chromosome is the CFTR gene on in *Bos taurus*? (Hint: Expand the ‘Animal Models’ box and click on ‘NCBI HomoloGene’ Information about this gene in other species is available on the right of the page).

Go back to the previous page.

Let’s have a look at the pathways CFTR is present within.

Click on cellular pathways and then on KEGG ansd look at the pathway for *Vibrio cholerae* infection.

I hope you get the idea behind these last couple of questions-there is a huge amount of information available on genes and many ways to explore them.

Part 2 Exploring the general genome server at the ncbi

In this part of the practical will explore the (huge) range of genomes available at the ncbi. Some are very incomplete, some are not. I suggest you return to this site in your own time to explore the information on those organisms you find most interesting.

Go to <https://www.ncbi.nlm.nih.gov/genome/>

Click on ‘browse by organism’ (on the left) and then click on eukaryotes. Explore the ‘filter’ options. Is your organism of interest present the database? Or you could type in your organism name. e.g *Oryctolagus cuniculus* (European rabbit) this is the genome in which we found the first endogenous lentivirus. You can blast the organism you have selected with your gene of interest to find genes/proteins within the genome. e.g. In your own time you could try blasting the EIAV sequence (an exogenous lentivirus) from the first practical against the rabbit genome. You may need to click on the assembly to do this

Question 9 How many mammalian sequences are in the genome server?

Question 10 Click on the ‘Size Mb’. What is a Mb? Why are the genomes of some species apparently so small?