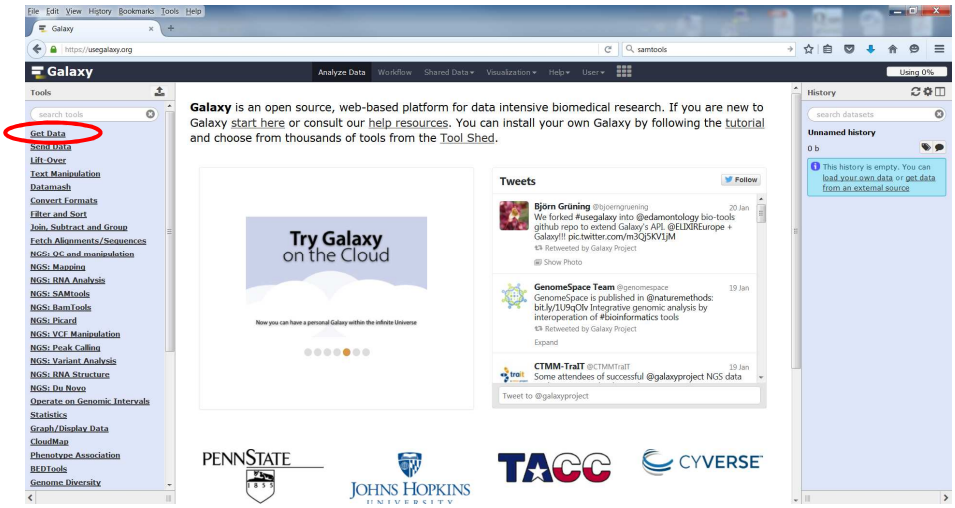
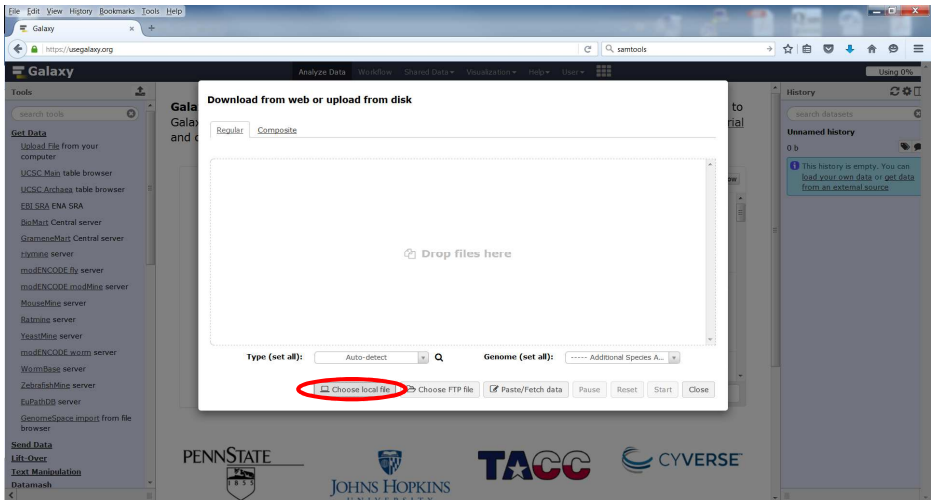
Open Galaxy and log in.

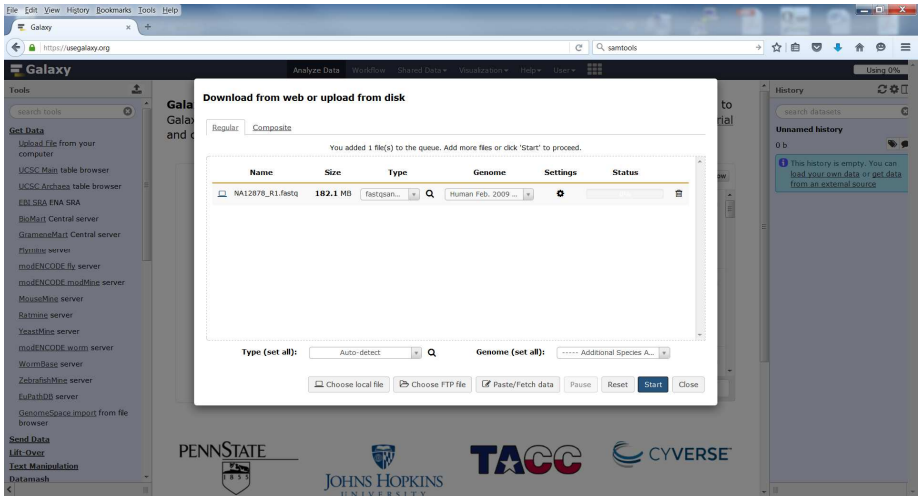
Click on **Get Data** in the left hand panel. Click on Upload File from your computer.



You will get a box like the one below. Select Choose local file.

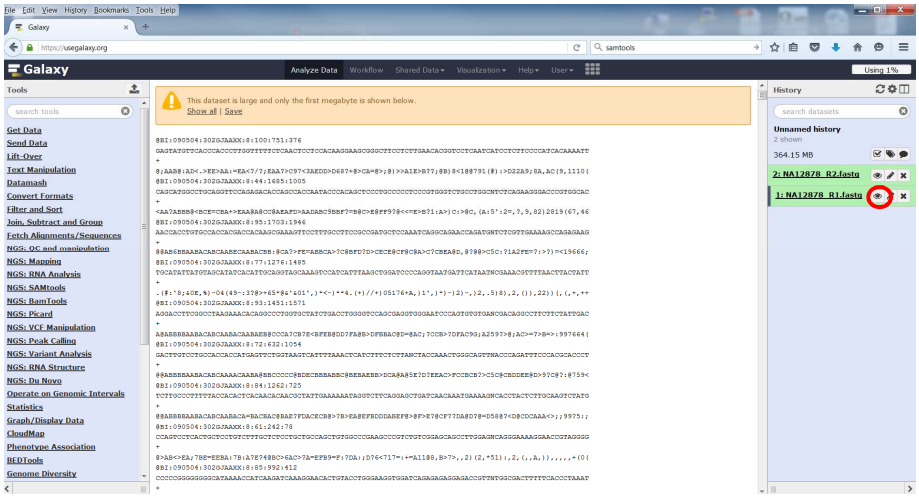


Scroll to where you saved the data files in the window that appears, select one the read 1 fastq files (labelled with ending R1.fastq) and click open. The file will appear in the download box as shown below. From the Type drop down menu select fastqsanger and from the Genome drop down menu select Human Feb. 2009 (GRCh37/hg19) (hg19). Click Choose local file again and repeat for the read 2 fastq (labelled ending R2.fastq).



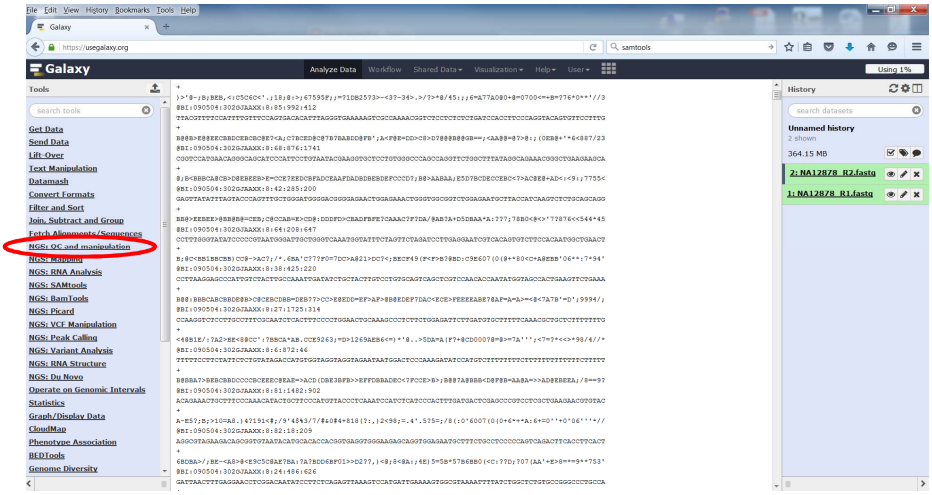
Then click Start. A green bar will appear in the Status column to show the uploading progress. The rows for both files will be shaded green when the upload is complete. This may take about 5 minutes. When the upload is complete click close to return to the main window. Both files will now appear in your history.

Open the read 1 file in the main window by clicking on the eye icon next to its name in the history panel. The result should look like the picture below.



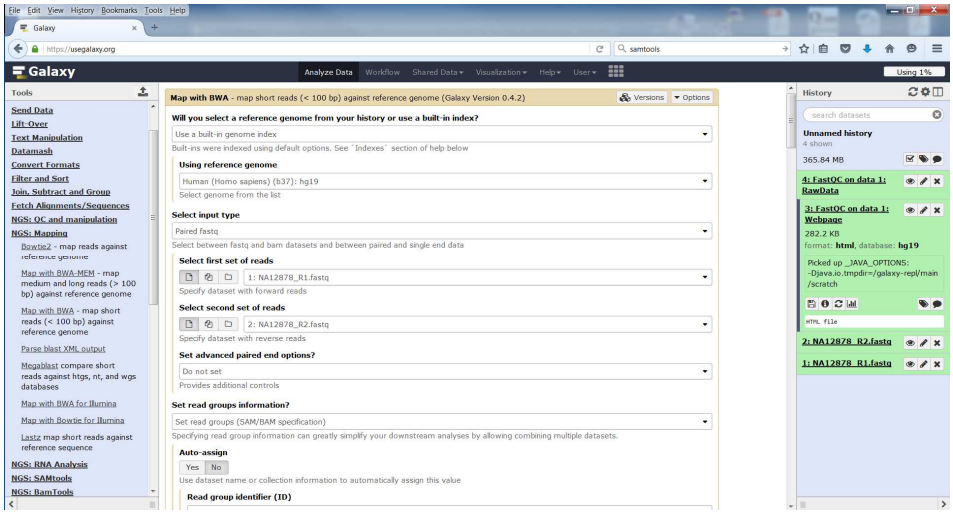
Look at the text. Can you see where the read identifier is? Which lines show the DNA sequence and where are the base qualities? Write down the first five bases and their Phred scores for the read labelled `@BI:090504:302GJAAXX:8:68:876:1741`. (you can scroll down the file with the slider on the right-hand side of the main window). Look in the read 2 file and find the read with the same identifier. What are the first five bases and their quality scores?

We will assess the overall quality of the sequencing using FastQC. Click on **NGS:QC and manipulation** in the left-hand panel and select FastQC.



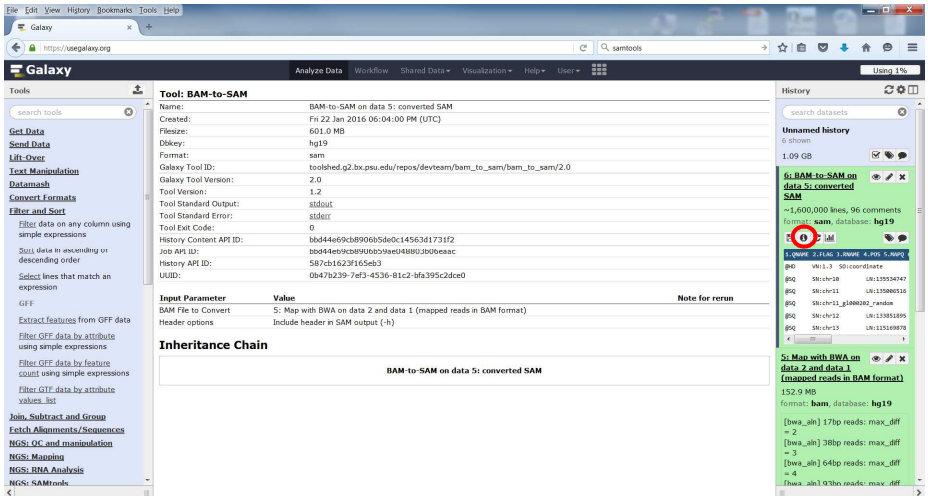
A series of options to supply and information about FastQC appears in the main panel. Read about the tool under **Purpose** and **Inputs and outputs**. Under **Short read data from your current history** select NA12878_R1.fastq from the drop down menu. Leave the other boxes with Nothing selected. Click Execute. Two new jobs will appear in your history. Jobs begin as grey with a clock symbol to show they have been queued. They turn yellow when they are running and green when they are finished. Look at the results labelled with Webpage. What is the read length? Under **Per base sequence quality** is a summary plot of the base qualities with position along the read. It is typical for the base quality to decline at the end of the reads, as is seen here. Scroll to **Per base sequence content**. We expect the percentage of different bases to be approximately the same.

Now we will align the reads to the human genome reference. Under tools choose **NGS:Mapping** and click on Map with BWA, this is the BWA-aln algorithm. Conveniently, it combines a number of commands that you would have to run if doing this in Linux, and produces an indexed bam file sorted by genomic coordinates. In the options keep Use a built-in genome index, under **Using reference genome** select Human (Homo sapiens) (b37): hg19, and keep Paired fastq as the **input type**. Under **Select first set of reads** choose NA12787_R1.fastq from the pull down menu and choose NA12787_R2.fastq under **Select second set of reads**. Do not set any advanced paired end options.



Under **Set read groups information?** Select Set read groups (SAM/BAM specification) and type in the following into the respective boxes: **Read group identifier (ID)** - BI_090504_302GJAAXX_NA12787_lane_8, **Read group sample name (SM)** - NA12787, **Library name (LB)** – NA12787_lib_1, **Platform unit (PU)** - 302GJAAXX_lane_8. Then click Execute.

BAM files are not-human readable. To convert the BAM into a readable SAM file select **NGS: SAMtools** from the tool menu and click on BAM-to-SAM. For **BAM File to Convert** choose the file created by the BWA mapping and leave the **Header options** as Include header in SAM output. Click Execute. You can see the file size by clicking on the file name in the history to expand the information and then clicking on the i icon.

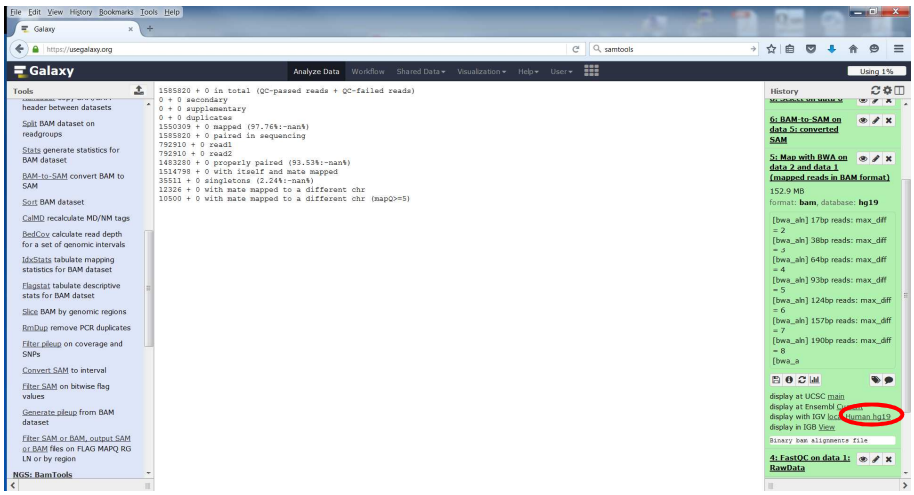


What size is the SAM file compared to the BAM file? Open the SAM file in the main window. The first lines, starting with @ tell you the contigs in the reference genome and their length (LN). We will see where the read that we looked at before has mapped to. To find the read, select the **Filter and Sort** tool and click on Select. Check that the correct file is selected, under **that** keep Matching and in the box under **the pattern** type 68:876:1741. Click

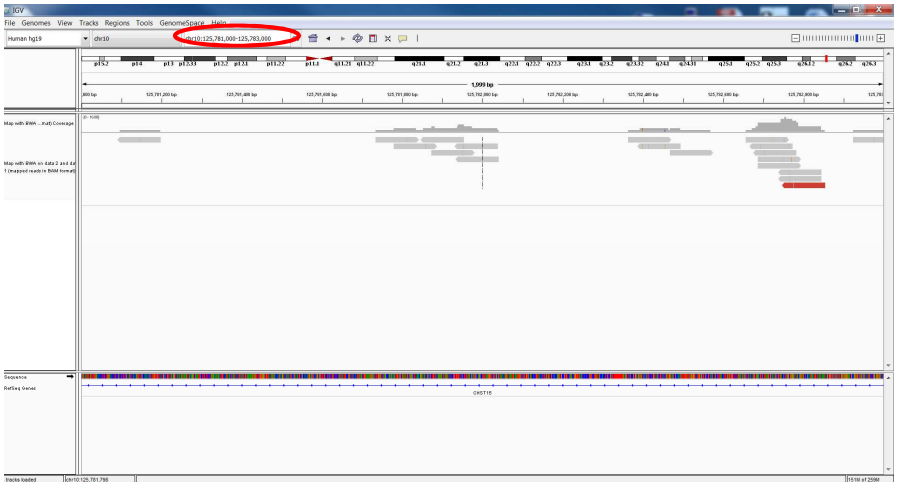
Execute. Two reads should be found. Which is read 1? Which mapped to the reverse strand? Which read mapped furthest 5'. Sketch a picture of how the reads lie on the two strands. Are their positions correct?

You can get a summary of the alignment with flagstat from SAMtools. From tools, select **NGS: SAMtools** and click on Flagstat. Choose the BAM file from the history and click Execute. Open the output. How many reads were there? What proportion were mapped?

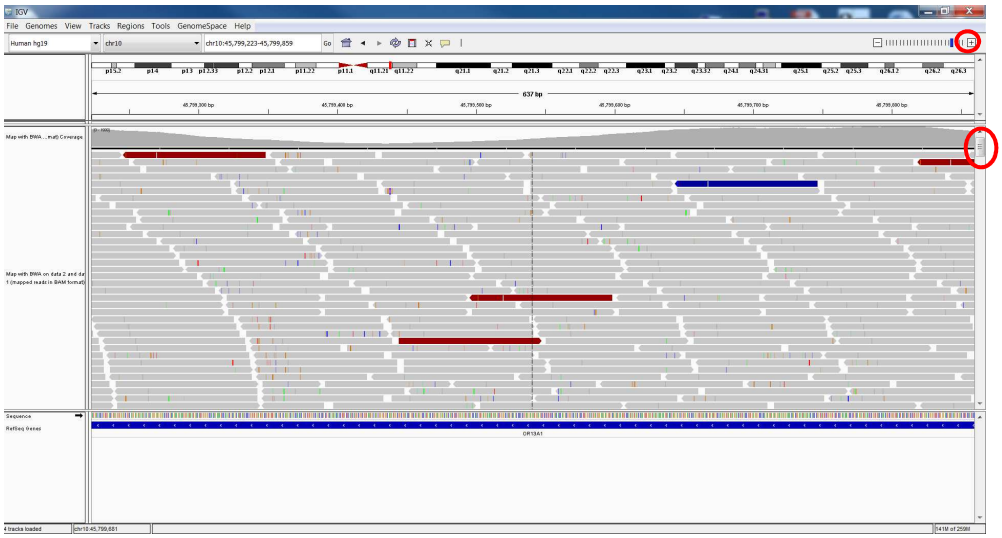
Galaxy lets you see your alignments using IGV (integrative genome viewer). Click on the BAM file in the history to open up the details and then click on display with IGV Human hg19.



You may get a pop up window asking if you want to open the file with Java. Click OK. If asked Do you want to run this application?, click Run. It may take a few minutes to load and will open in a new window. Human hg19 should appear in the top left window. The BAM file may not load initially. If you do not see the text Map with BWA in the left column, return to Galaxy and this time select IGV local in the details for the BAM file. The chromosomes appear in the top bar of the IGV window. In the box in front of Go, type chr10:125781000-125783000 and click Go. The window should change to look like below.



The red bar on the chromosome picture at the top shows you where you are and the genomic coordinates of region illustrated are given. The first row under the coordinates shows plots of the coverage (depth). Underneath are representations of the reads. The point indicates the direction of sequencing (if the point is to the left the read was sequenced on the reverse strand). If you click on a read a pop up box gives the read name and details about it. Find the two reads from cluster 68:876:1741. Does their positioning match with the sketch you drew? You can go to the position of a particular gene. Type OR13A1 in the position box and click Go. Click on the plus box at the top right to zoom in. Reads shaded dark red have larger than expected insert sizes and reads shaded dark blue have smaller than expected insert sizes. Bases where a read differs from the reference are coloured according to the changed base and are darker the higher the quality score of that base. You can move along the gene by scrolling in the coverage row. Not all the reads are shown. Use the bar on the right hand side to scroll down to see more.



Close the IGV window when you have finished.

These data are from a targeted sequencing experiment aimed at certain genes and it would be useful to know how many reads there are for each gene. To this we will upload the file called target_regions.bed into Galaxy. This file contains the start and end positions (in bed format which means that the start is -1 of the genomic position) of the genes. This time, select bed under **Type** and again choose Human Feb. 2009 (GRCh37/hg19) (hg19) for the **Genome**. To find the coverage choose **BEDTools** and select Compute both the depth and breadth of coverage. Under **Count how many intervals in this BED/VCF/GFF/BAM file** choose the BAM file from mapping with BWA and under the **overlap the intervals in this BED file** choose the uploaded target regions file. Click Execute. Look at the results. If you scroll across, after the details of each interval, column 5 gives the number of reads that overlapped the region. (column 6 tells you how many bases in the region were covered by at least one read, column 7 the total number of bases in the region and column 8 the percentage of bases covered by at least one read). Which region had the most reads? Which region had the highest percentage of bases covered?

Finally, we will look at how to detect duplicates in our read data. It is important that duplicates are filtered out because an assumption of the data analysis is that the reads are a random selection from the genome. Select **NGS: Picard** and choose MarkDuplicates. Choose the BAM file. In the box under **Regular expression that can be used to parse read names**… type NULL. In the box under **Select validation stringency** choose Silent. Click on Execute. Picard creates a new BAM file dataset in which the duplicates are flagged and can therefore be ignored by some software. Run SAMtools Flagstat on the new BAM file to see how many duplicates were identified.