Overview

Analysis of NGS data involves three steps: alignment; variant calling; and annotation, and it is essential to assess data quality and performance at each step. At the end of this exercise you will be able to:

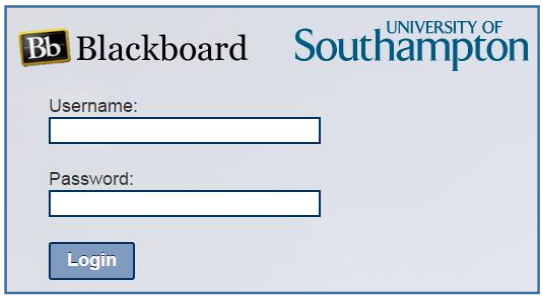
1. Use the Galaxy suite of bioinformatic tools
2. Assess the quality of raw NGS data in fastq format prior to alignment
3. Align NGS data to the reference human genome using BWA
4. Describe the contents of Fastq, SAM and BAM files
5. Make an assessment of the alignment process
6. Conduct quality control filtering of reads
7. Visualise aligned data

Trial data

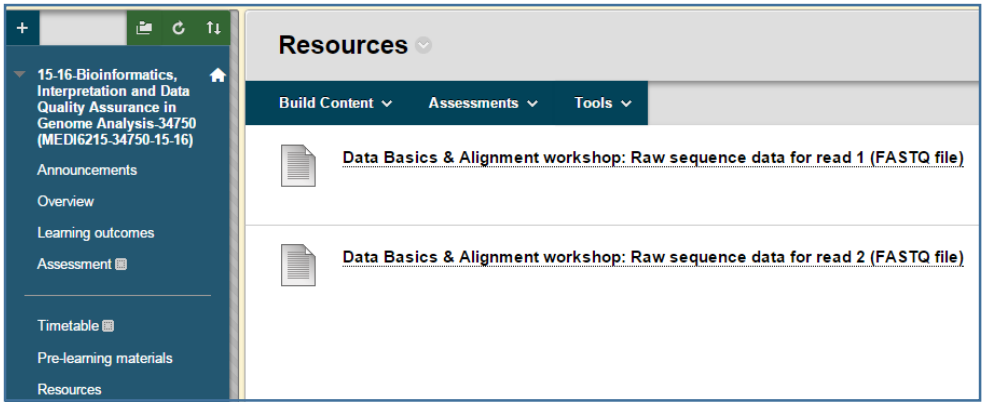
The sequence data that you will be analysing is from a 25-year-old male who presented with hearing loss in the left ear and some deterioration in visual acuity especially at night. He had also noticed some numbness of his left arm and difficulty in putting on a jumper due to some weakness of his left shoulder. He has no relevant family history. An MRI scan showed left acoustic neuromas, a mass under his left scapula and a mass impinging on his left brachial plexus.
The patients exome was sequenced using the paired-end method on an Illumina HiSeq 2000 following target enrichment by Agilent SureSelect. Your aim over the next three practicals is to analyse this data and determine if there are any disease causing mutations, and if so what disease is implicated. While following the tasks below, think about the genes that should be prioritised for mutation screening given the patients symptoms.

Let's begin

1. Login to blackboard: https://blackboard.soton.ac.uk/



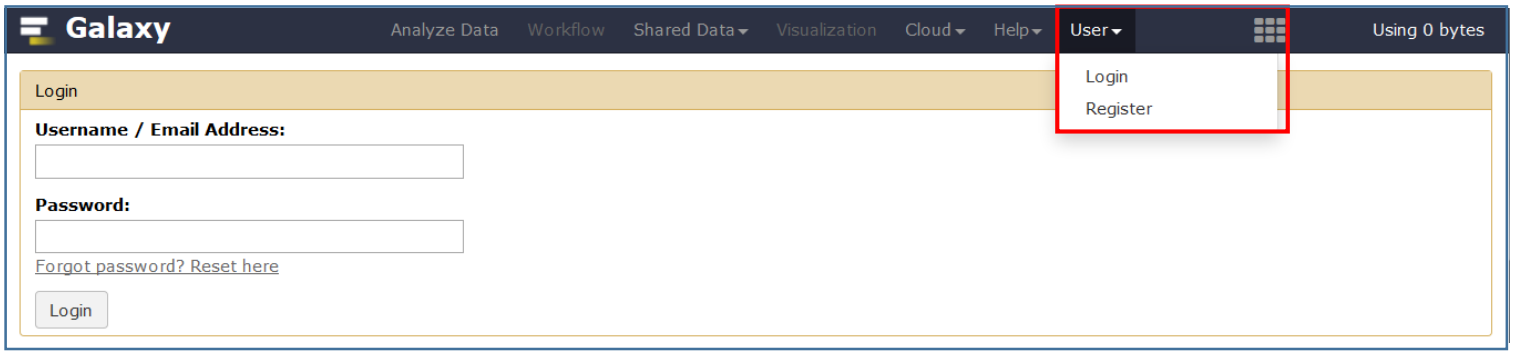
2. Navigate to the course resources and download the raw sequence data (two files) and a file describing the sequenced region to your computer.



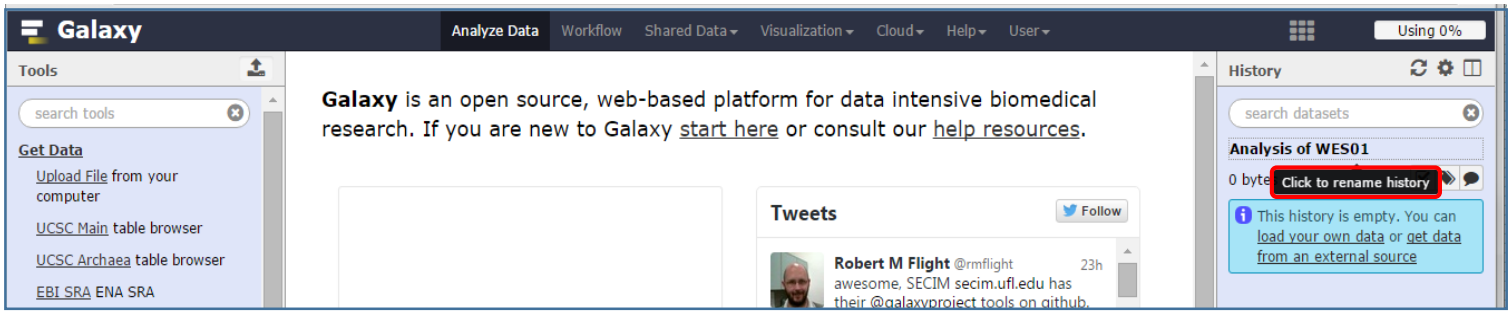
3. Go to https://usegalaxy.org/

The available "Tools" are in the left tool pane, the central panel is your working area where you can visualise your data, select tool options and execute jobs. Your "History" is shown in the right hand panel. Here you will see your data, results and a history of all the tools that you have used.

4. Register for a galaxy account and login

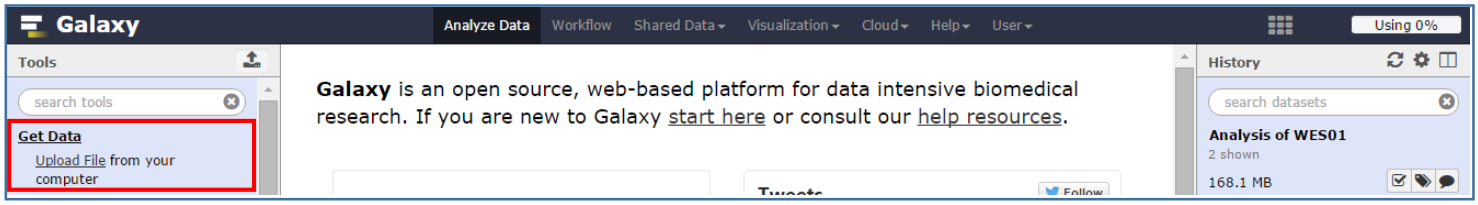


5. Rename the history for this session (Click on name, type new name, press return)

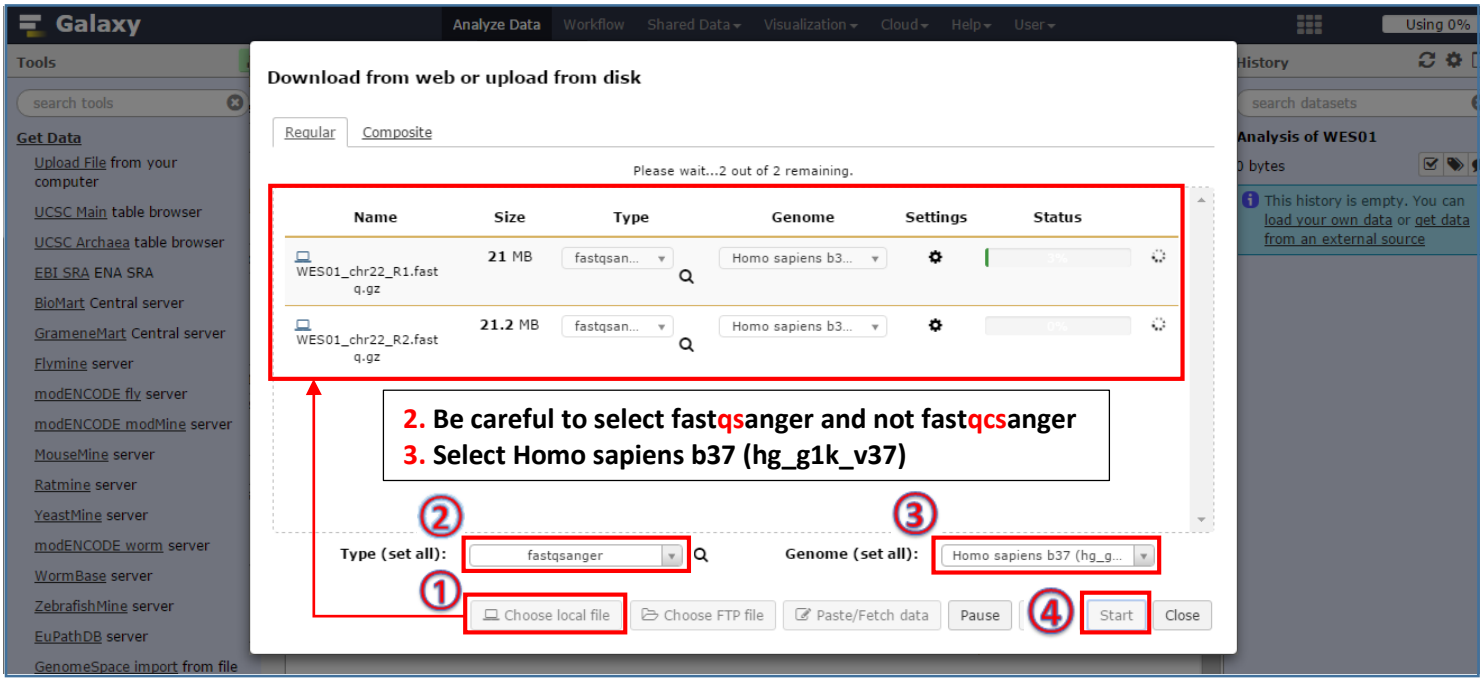


Uploading Files

1. In **Tool Pane**: Go to **Get Data** > Upload File from your computer

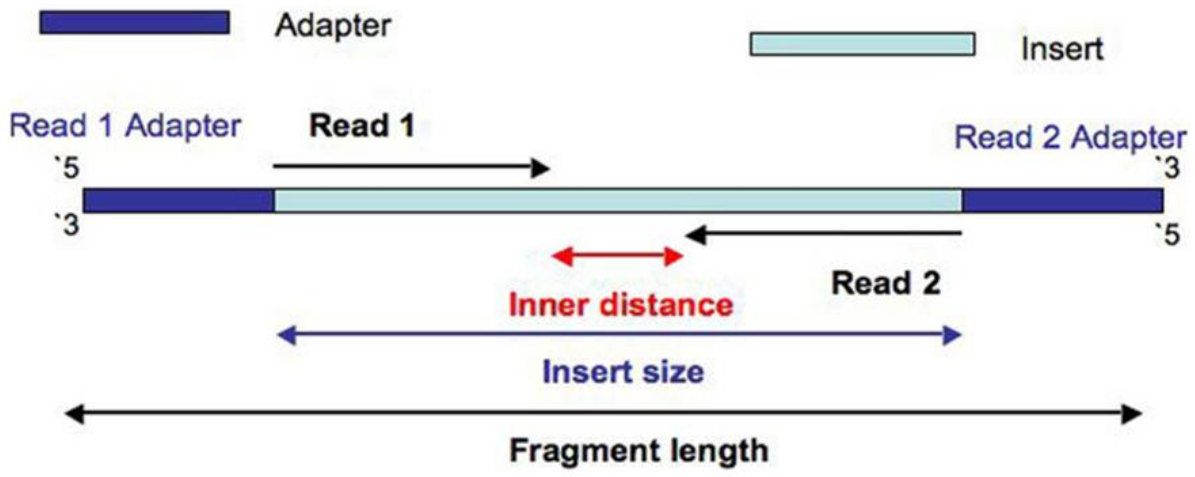


2. Click Upload File

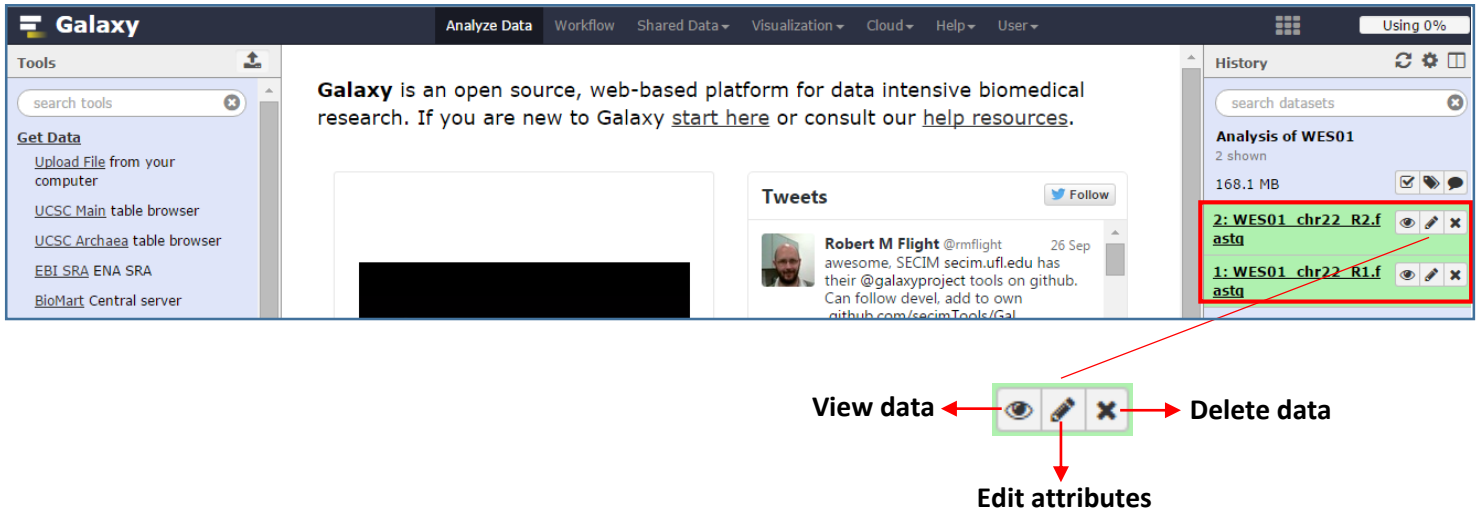


One lane of paired end sequencing was performed so you have two files of raw sequence data (WES01_chr22_**R1**.fastq.gz and WES01_chr22_**R2**.fastq.gz) which contain all the sequence data for read 1 (R1) and read 2 (R2) respectively (Figure 1). To save on computing time and disk space, the NGS data for WES01 has been filtered to contain reads mapping to chromosome 22 only and the files have been compressed (hence the .gz extension).

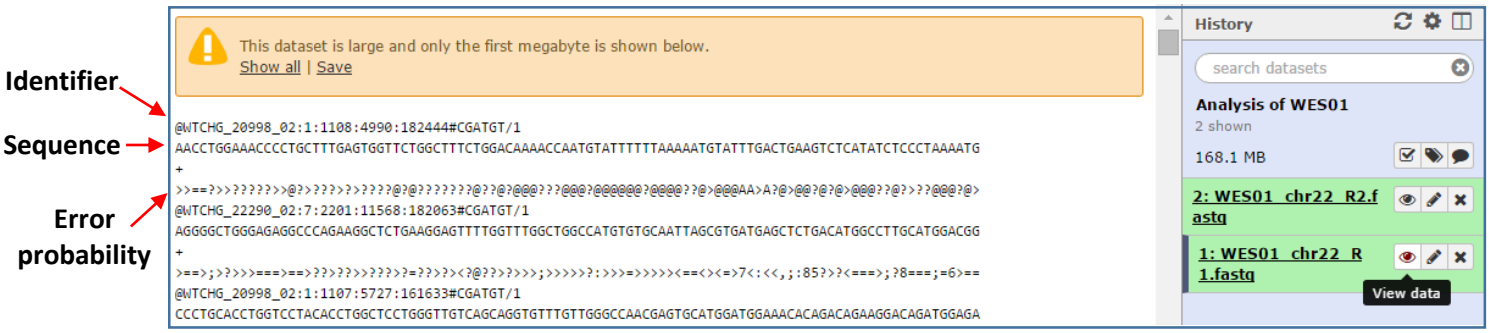
Figure 1. Paired end sequence data



3. The uploaded files will appear in the **History Pane**.



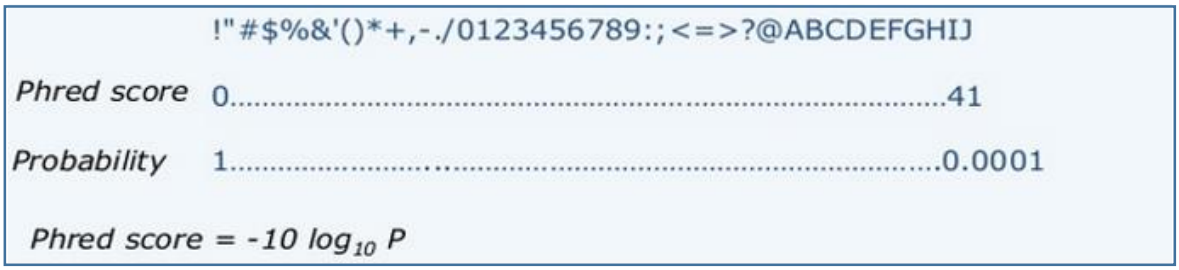
4. View the fastq file for read 1 by clicking the view data button  in the **History Pane**.



The raw NGS data is held in fastq format as described in lecture 1 and in more detail here: http://en.wikipedia.org/wiki/FASTQ_format. Fastq files are the simplest and most generic way of storing read sequences and qualities. Each read has 4 lines of data; an identifier, the sequence, + and the sequence error probabilities.

It is more efficient to store the sequence quality scores as characters because they require less disk space than numbers (i.e. 1 character versus multiple digits). Characters are also more reliable and portable than numbers. Sequence quality scores on the phred scale are determined by mapping the ASCII character to its associated number (Figure 2).

Figure 2. Phred quality scores

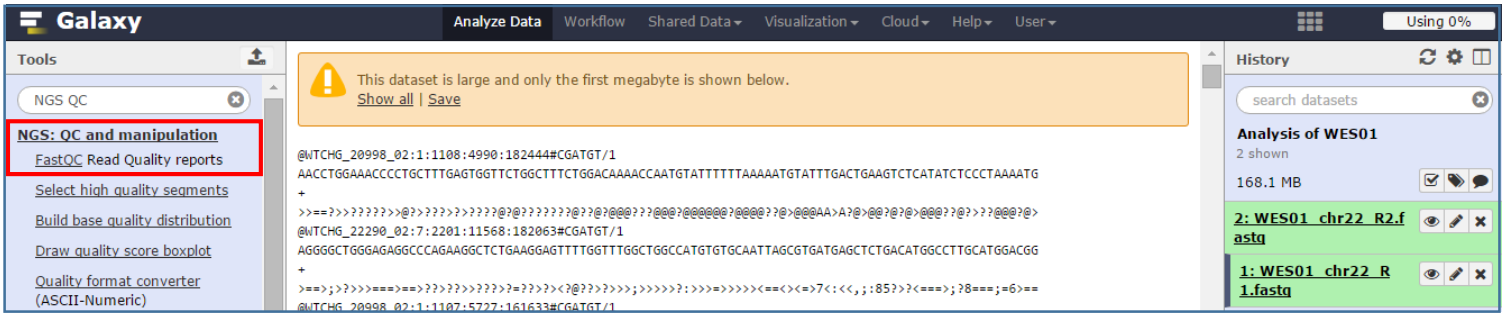


Q1. Use this website (http://grand-prismatic.blogspot.co.uk/2013/02/fastq-quality-score-convesion-table.html) to determine the error probability of the first base pair of read 1.

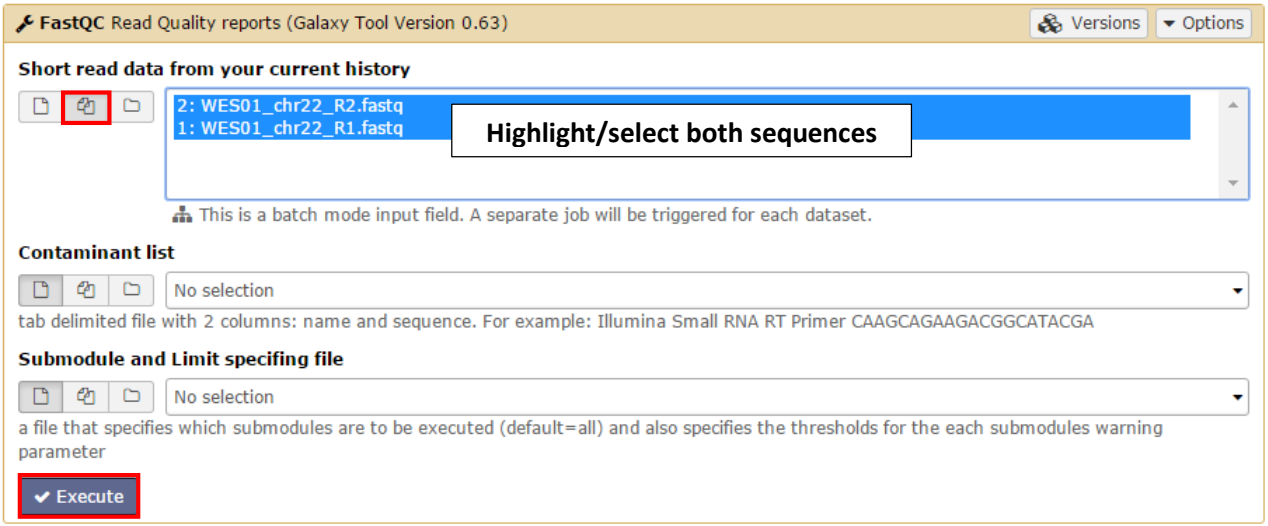
Assess the quality of raw sequence data

To assess the quality of the raw sequence data and to guide quality control we will use a program called FastQC. The program outputs summary graphs and tables that show if there are any problem areas, which could influence assembly or variant calling if not addressed. You can learn more about the program here (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

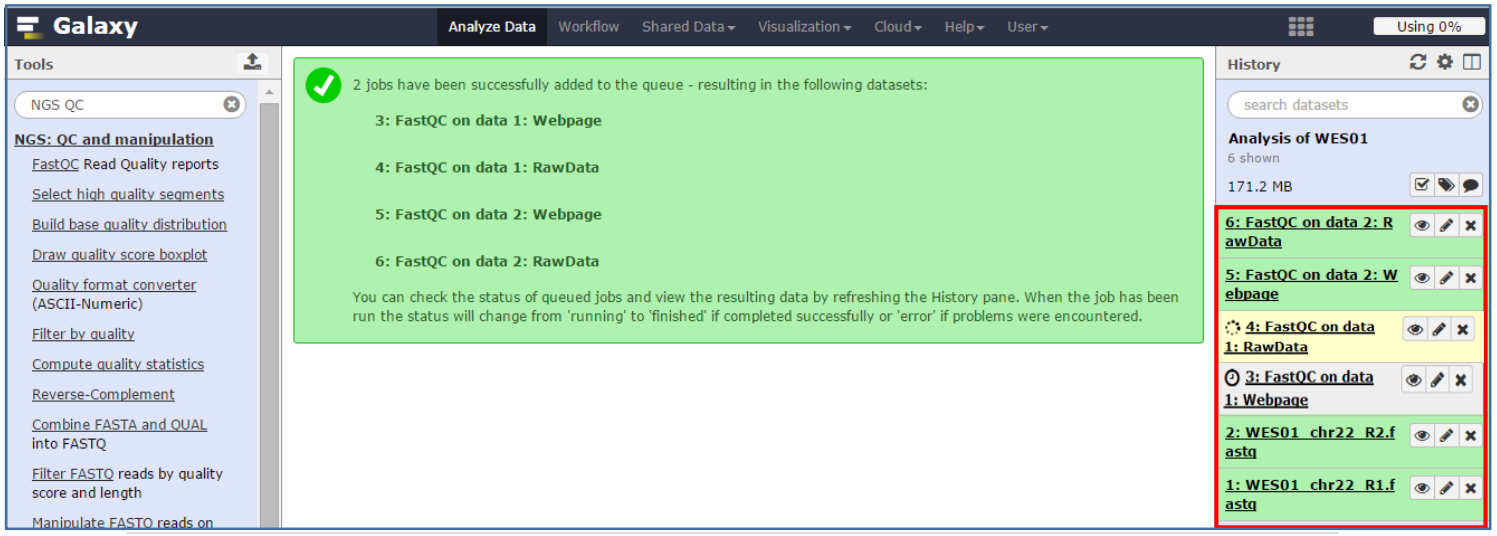
1. In **Tool Pane:** Go to **NGS: QC and manipulation** > FastQC Read Quality reports



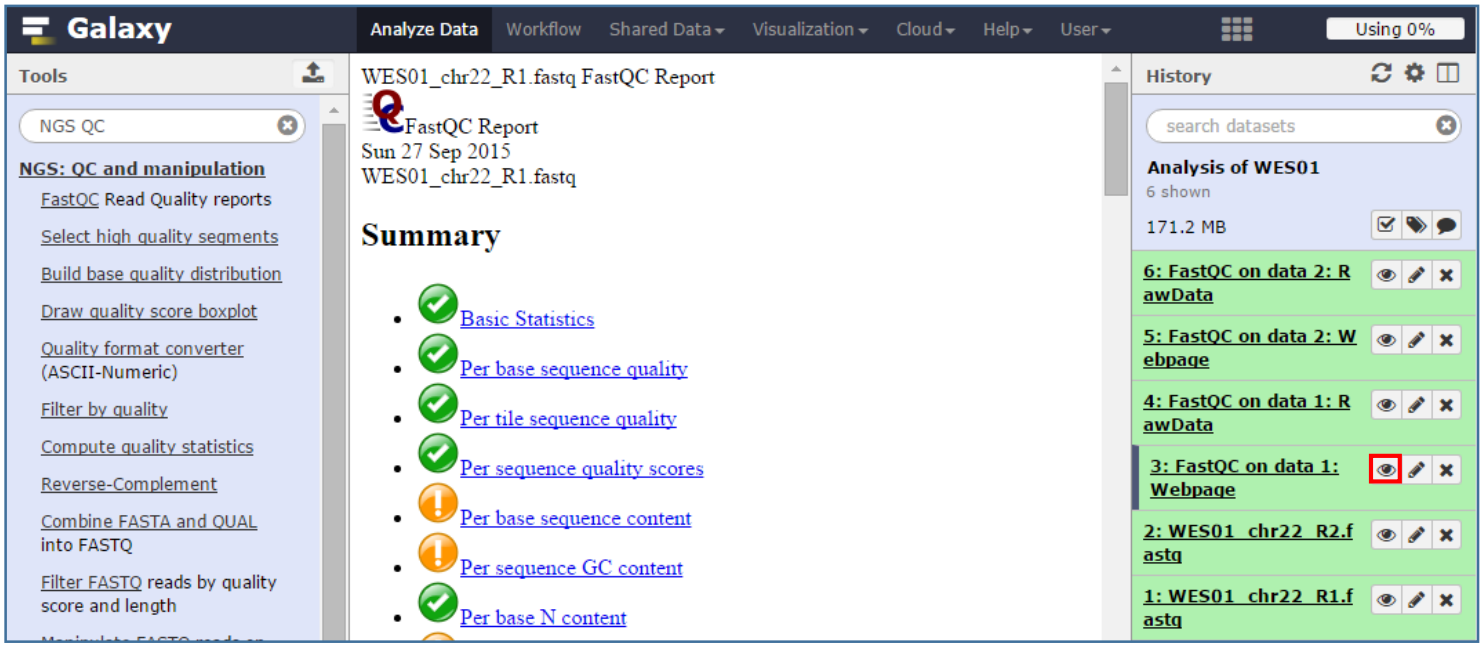
2. Select multiple datasets, highlight both fastq files and click Execute



The FastQC reports will be shown in the **History Pane.** Queued jobs in grey with a clock symbol, running jobs in yellow with a buffering symbol, finished jobs in green, failed jobs in red with a cross.

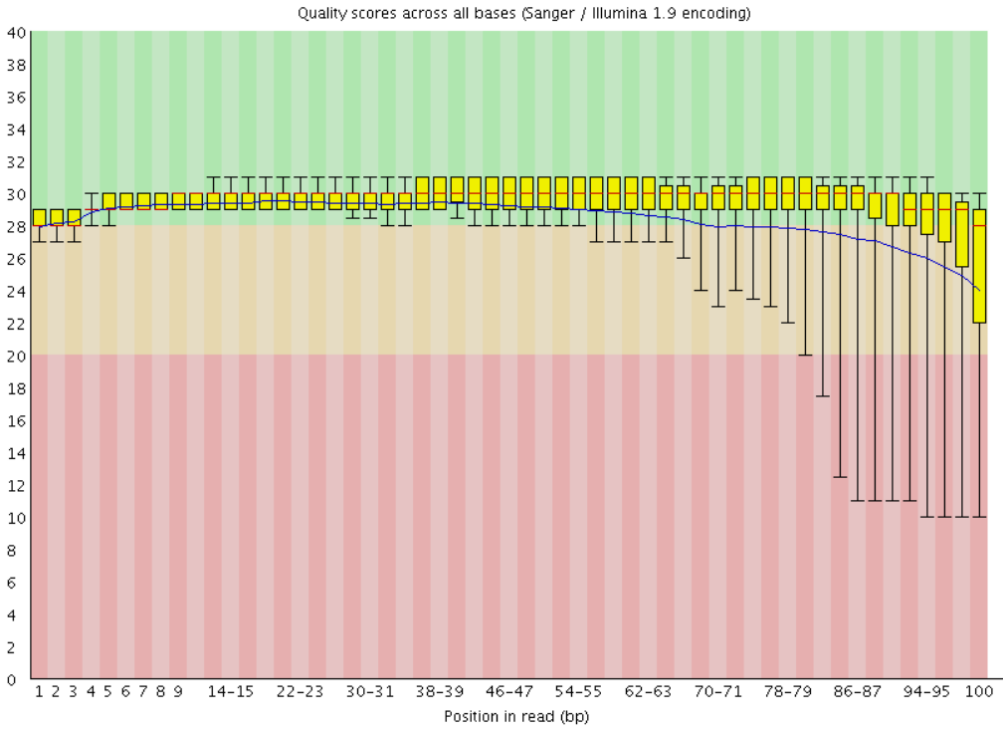


3. Click **View data** button for FastQC on data 1:Webpage to view the report



Use the FastQC reports on data 1 and 2 to answer:
Q2. How many reads do the files contain?
Q3. How long are the reads (bp)?
Q4. Has either file failed any of the sequence quality checks?

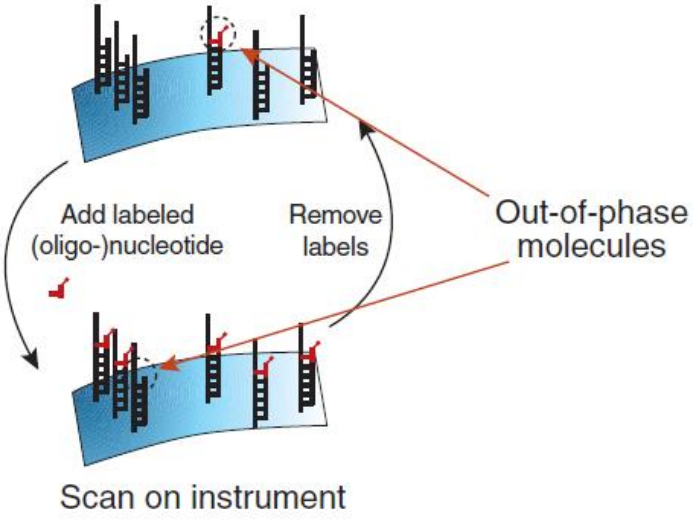
Figure 3. Per base sequence quality for WES01_chr22_R1.fastq



Looking at the per base sequence quality you will notice that the average base quality drops towards the end of reads (Figure 3). This drop in quality is typical for ensemble-based sequencing by synthesis (SBS) methods, such as Illumina, which add complimentary bases one at a time in a cluster of identical

sequences to determine a consensus sequence from the 'average' sequence signal over all copies in the cluster. As nucleotides are added some of the sequences in a cluster grow at a different rate and become desynchronized which reduces the accuracy of the 'average' sequence signal (Figure 4).

Figure 4. Read-length and phasing (From Fuller et al. 2009: Nat Biotechnol. doi: 10.1038/nbt)

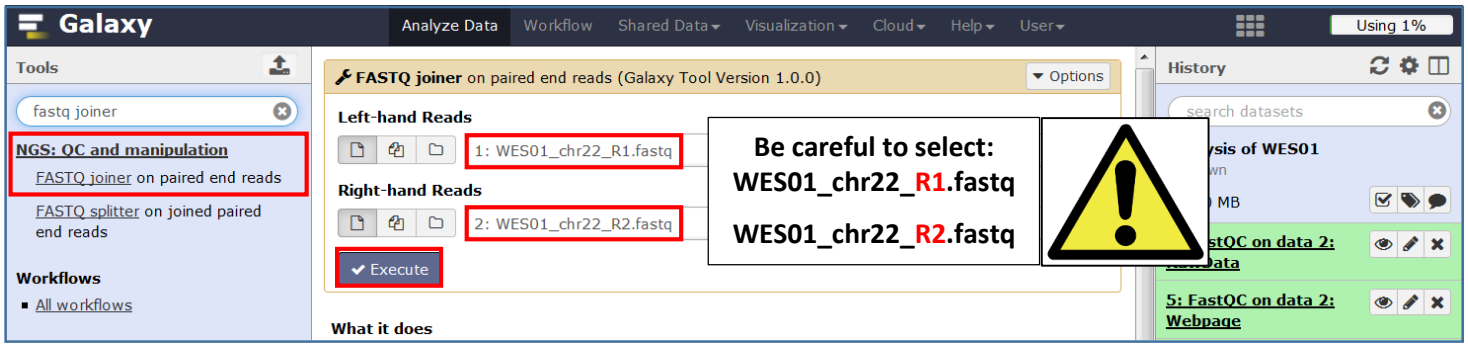


Another particularly important plot is that of 'Overrepresented sequences', which lists sequences that account for more than 0.1% of the total. The presence of an overrepresented sequence suggests that the sequence is biologically significant, or that the library is contaminated or has low diversity. To check for contamination, each overrepresented sequence is compared to a database of common contaminants such as sequencing adaptors which can then be removed from the raw FastQ data.

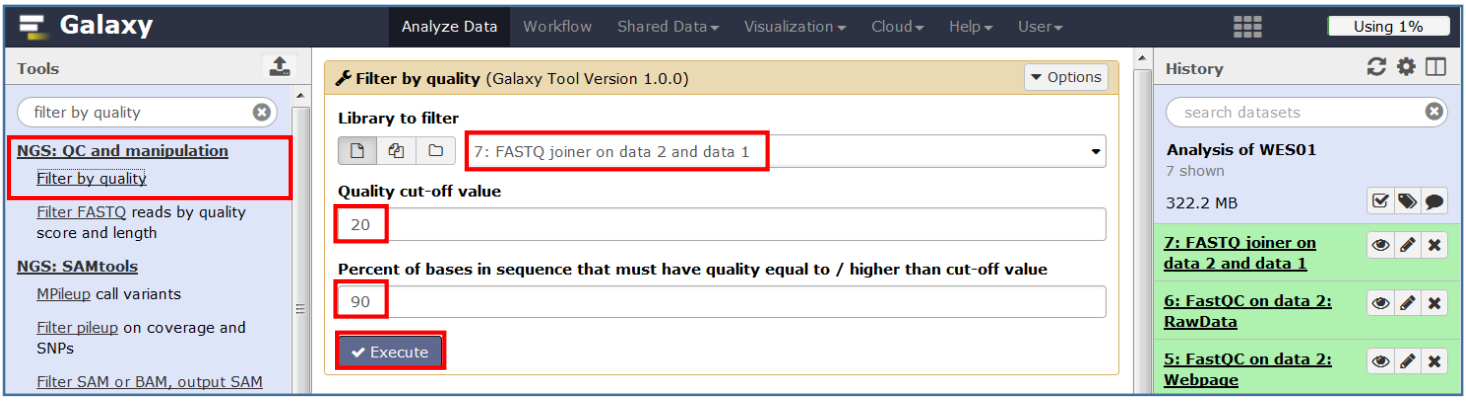
Filter reads based on quality

In a typical analysis you may want to raise technical issues identified by FastQC such as low read count, poor quality, and overrepresented sequences with the data provider. To ensure that only data of a certain quality is used for further analysis we will exclude low quality reads. In paired-end data there are two fastq files per lane of sequencing which are synchronised so that matching pairs are stored in the same line of each file (eg the read in line 1, file 1 is paired with the read in line 1, file 2 and so on). To maintain this order when filtering, the fastq files need to be joined and the reads have to be removed as a pair.

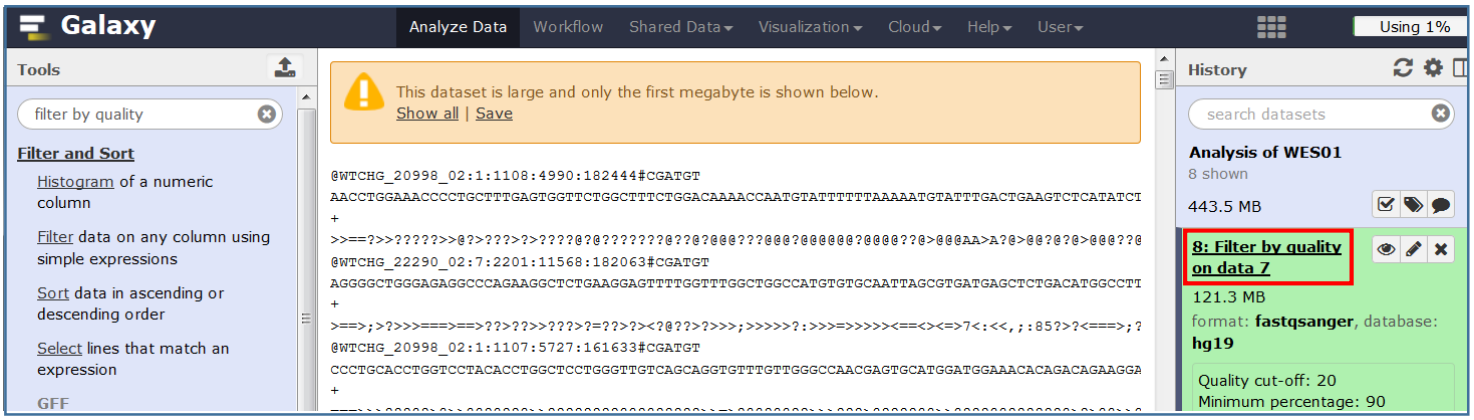
1. In **Tool Pane**: Go to **NGS: QC and manipulation** > FASTQ joiner



2. In **Tool Pane**: Go to **NGS: QC and manipulation** > Filter by quality
These setting will keep reads with a Phred quality score of 20 or more for 90% of its bases.



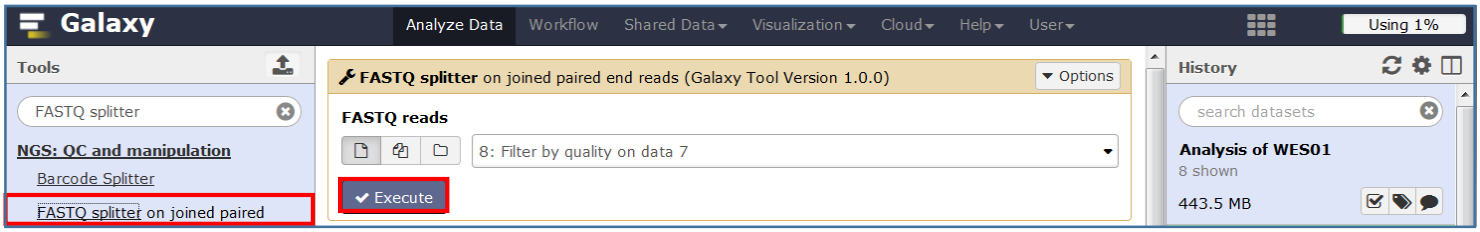
3. Clink the link (**8: Filter by quality on data 7**) to get details on the number of reads removed



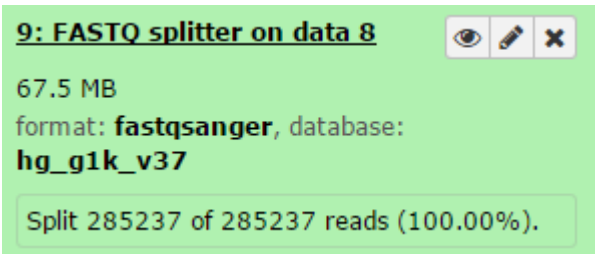
Q5. What number and percentage of reads were removed?

Now split the filtered file back into two fastq files ready for mapping.

4. In **Tool Pane**: Go to **NGS: QC and manipulation** > FASTQ splitter

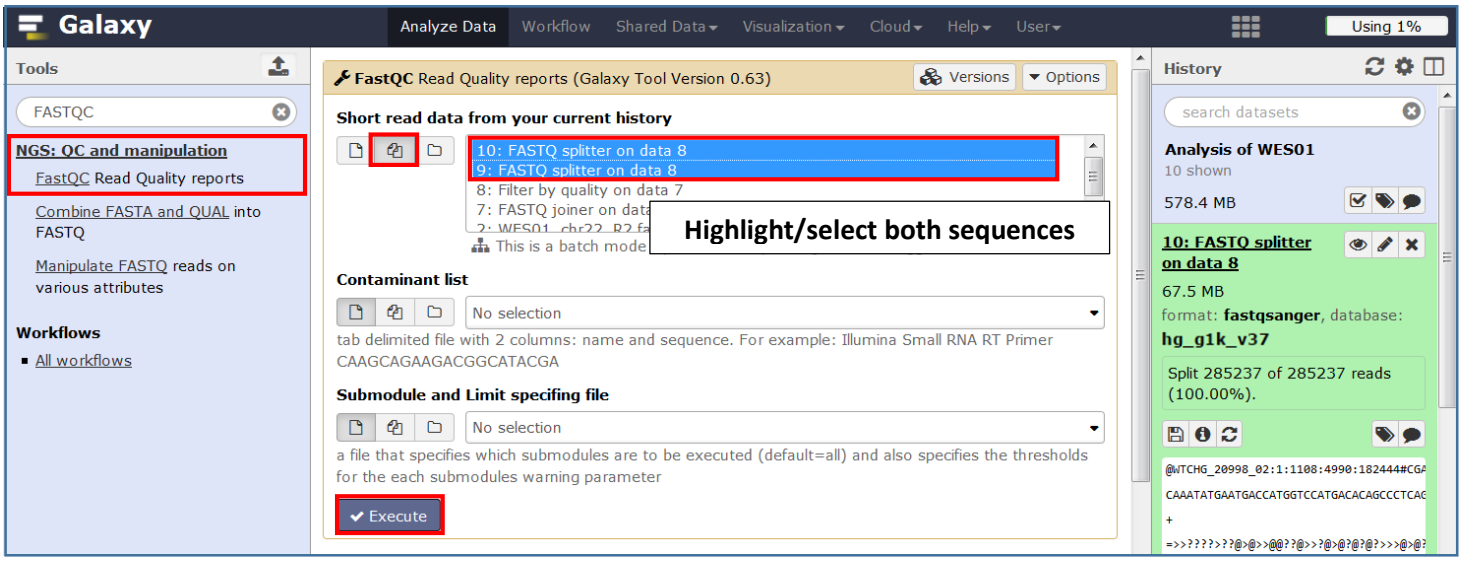


Check that the files have been successfully split with the same number of reads.



Rerun FastQC on the filtered reads to see the effect of filtering.

5. In **Tool Pane**: Go to **NGS: QC and manipulation** > FastQC Read Quality reports



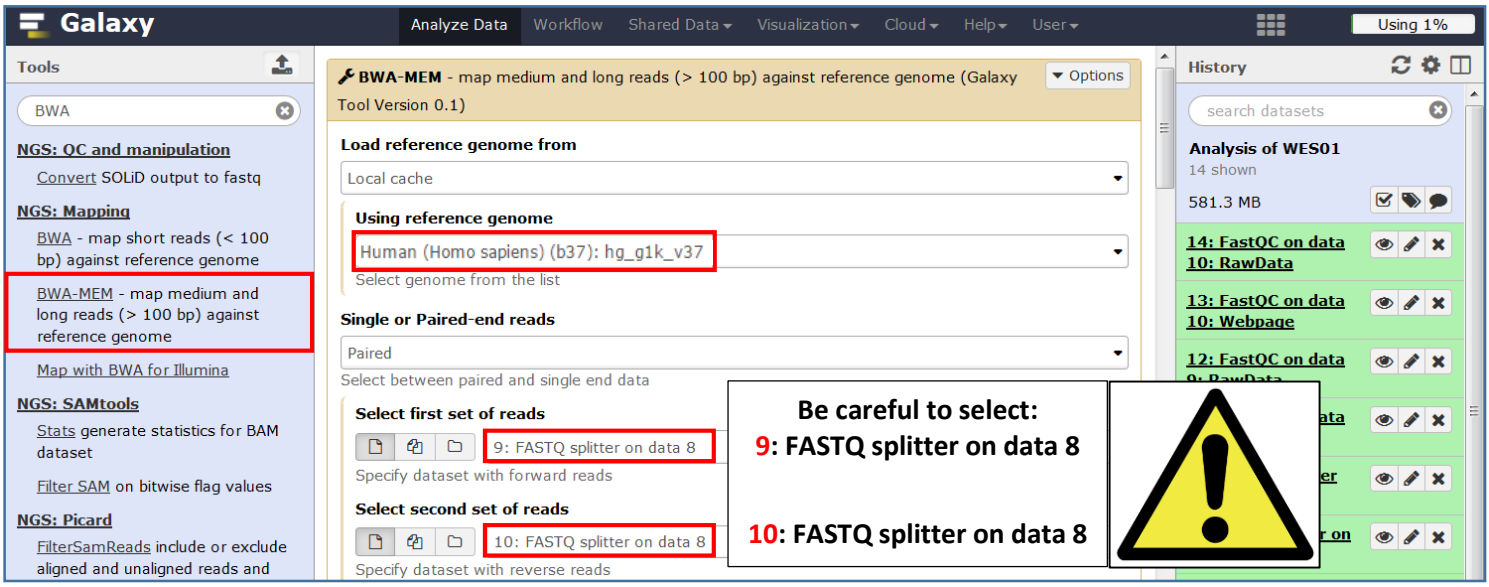
There are other ways to clean up reads such as trimming or filtering by other criteria that could be used depending on the data quality and requirements.

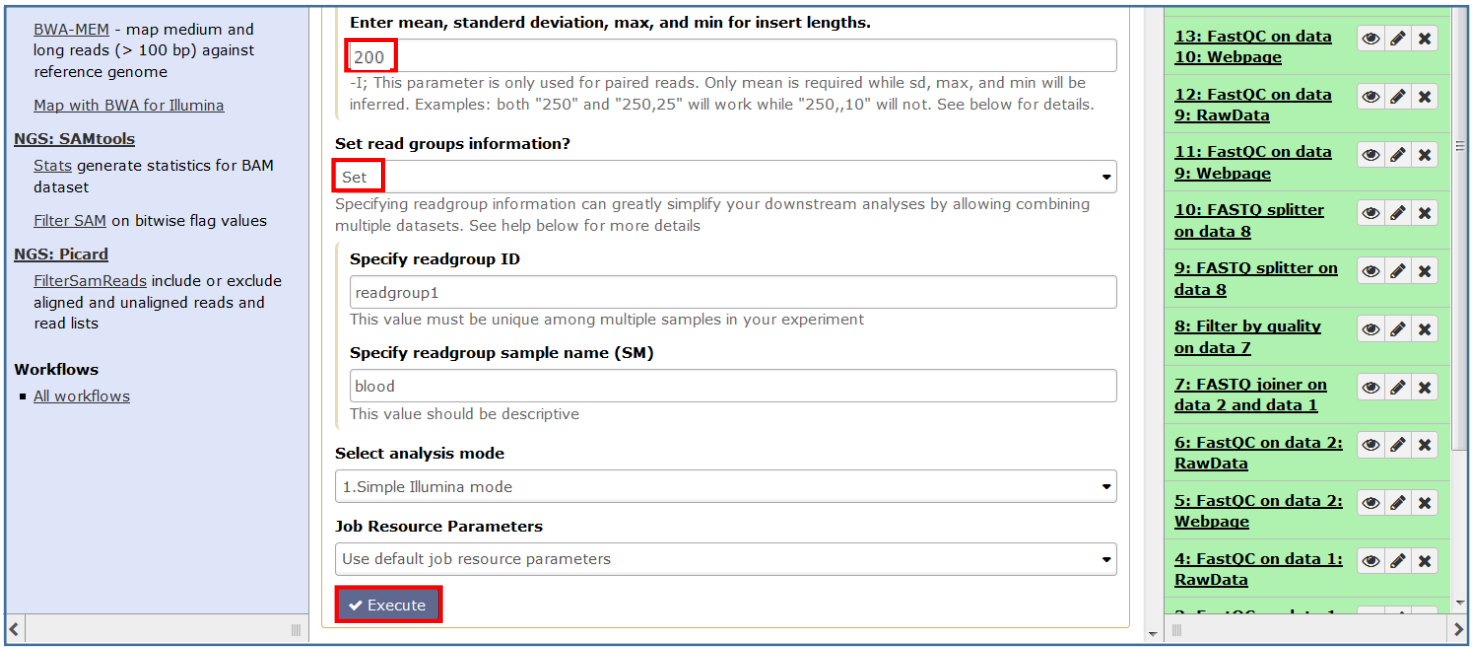
Align sequence data to the reference human genome

We will use a program called BWA-MEM (Burrows-Wheeler Alignment) to align the sequence data to the human genome. You can learn more about the bwa program here (http://bio-bwa.sourceforge.net/).

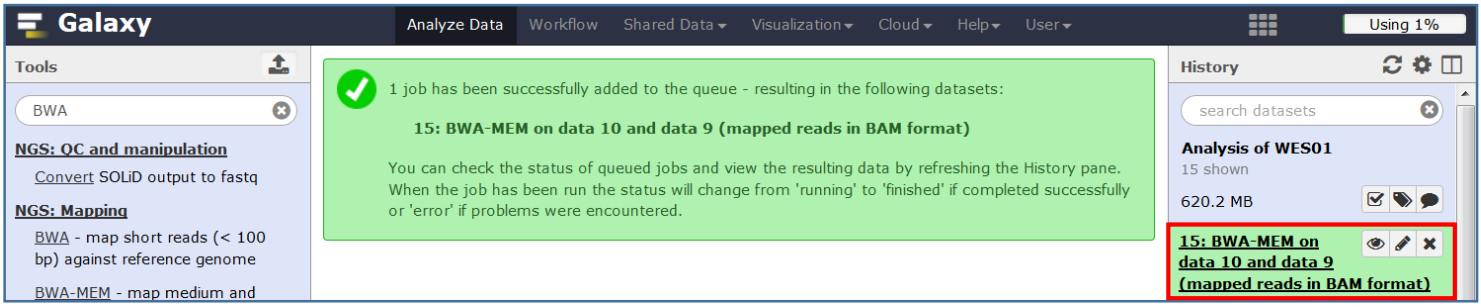
1. In **Tool Pane**: Go to **NGS:Mapping** > BWA-MEM

Select reference genome (hg_g1k_v37), first (9: FASTQ splitter on data 8) and second (10: FASTQ splitter on data 8) set of reads, enter a default insert length of 200bp (see Figure 1), set read group information, enter read group ID and sample name. Now click execute.





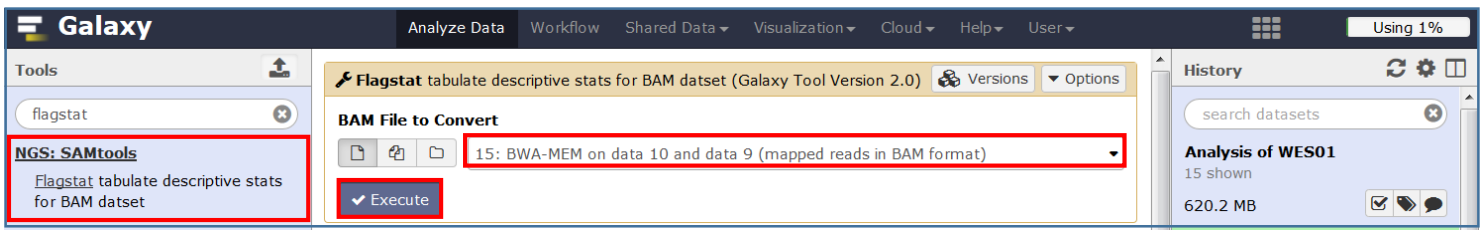
The alignment process maps the read data to the reference human genome and creates a Binary Alignment/Map file or BAM for short. The binary BAM file is not directly viewable and clicking the view button will download the file.



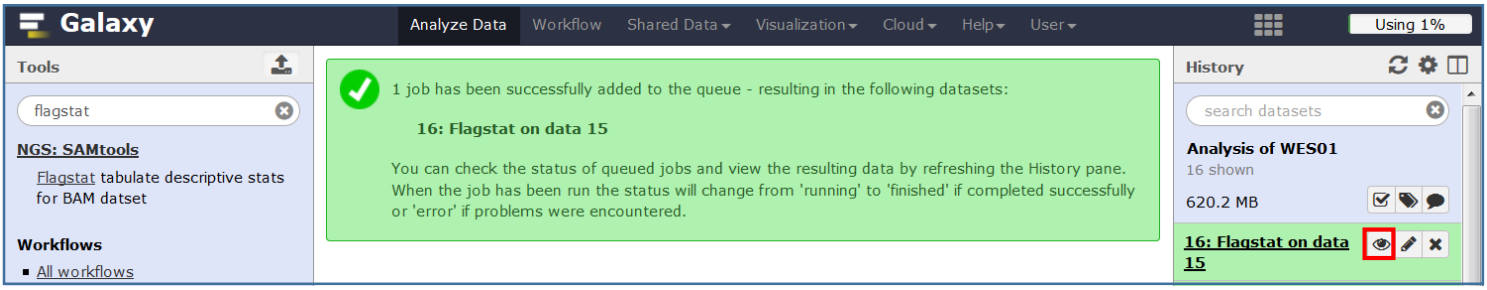
Generate alignment statistics

When aligning reads to the reference genome anywhere between 0 to 20% of reads are not aligned due to sequencing errors, sample contamination (eg bacterial or viral DNA), gaps in the reference genome and genome variation. Use SAMtools Flagstat to determine how many reads have been aligned.

1. In **Tool Pane**: Go to **NGS SAMtools** > Flagstat tabulate descriptive stats for BAM dataset



2. Click view data to look at the alignment stats for the BAM file



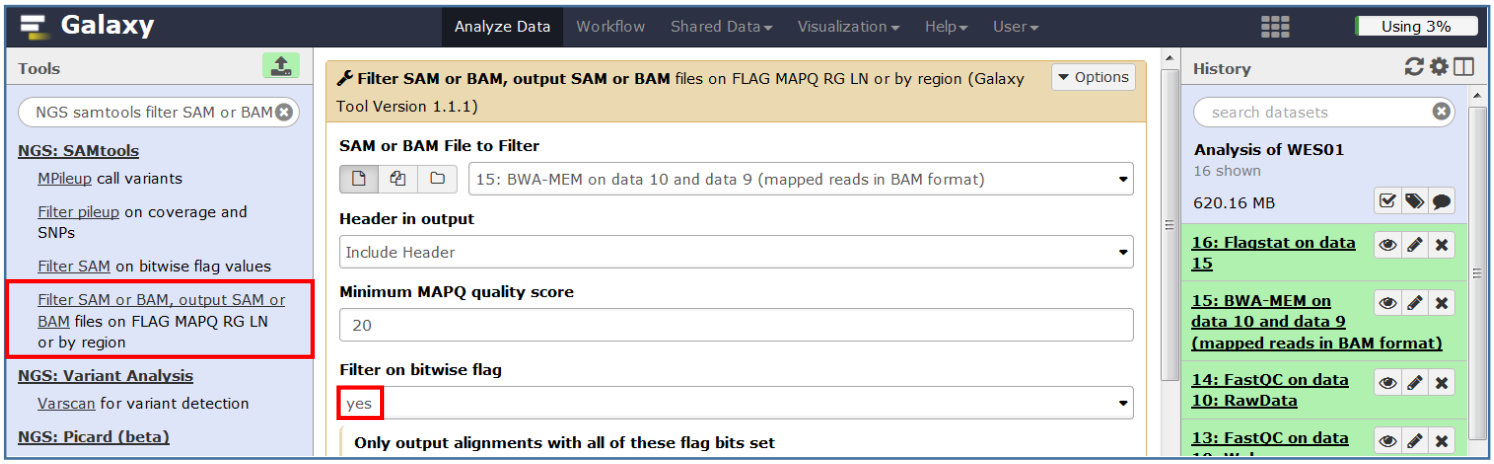
Q6. Use the Flagstat output to determine the percentage of mapped reads

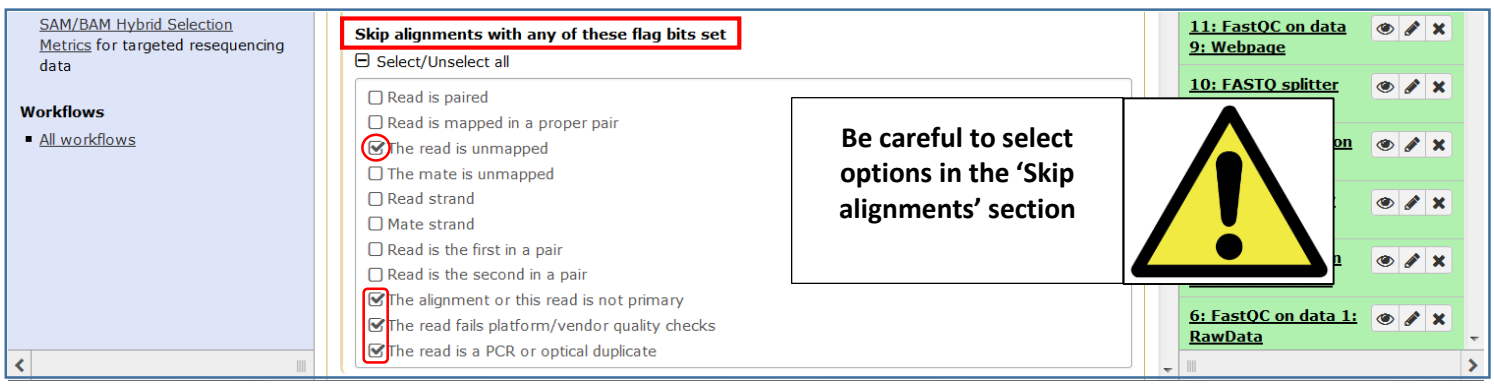
Filter BAM file

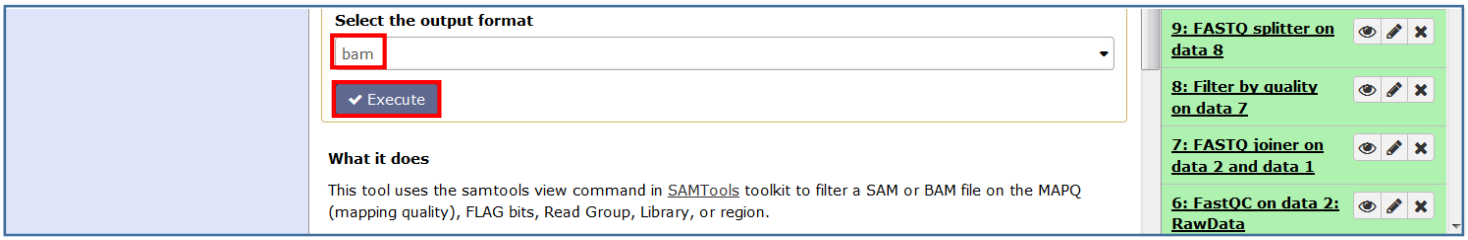
For variant calling, reads with low mapping quality (phred <20), unmapped reads, secondary alignments, reads failing platform/vendor quality checks, and duplicate reads that have the same start and stop position are typically removed or ignored because they can influence genotyping accuracy. For example, at heterozygous sites the two alleles should be evenly distributed (50% of reads have the A allele and 50% have the B allele) but if reads with the A allele are duplicated the A allele will become overrepresented and the site might be misinterpreted as homozygous for the A allele.

Use Filter SAM or BAM to make a new bam file which excludes these types of reads.

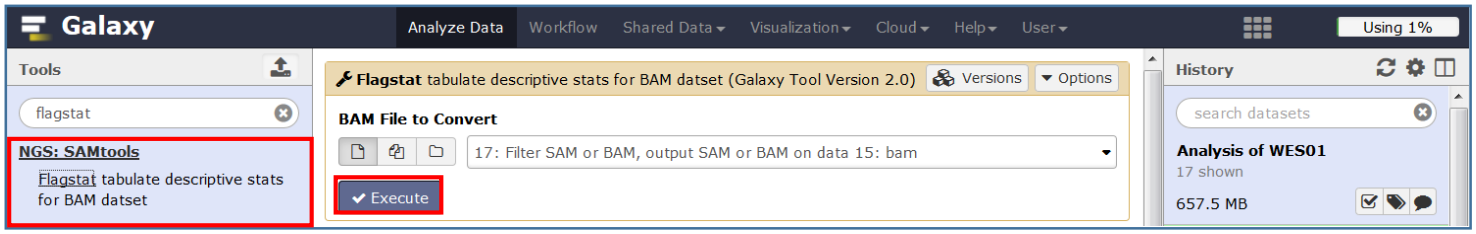
1. In **Tool Pane**: Go to **NGS: SAMtools** > Filter SAM or BAM, output SAM or BAM files



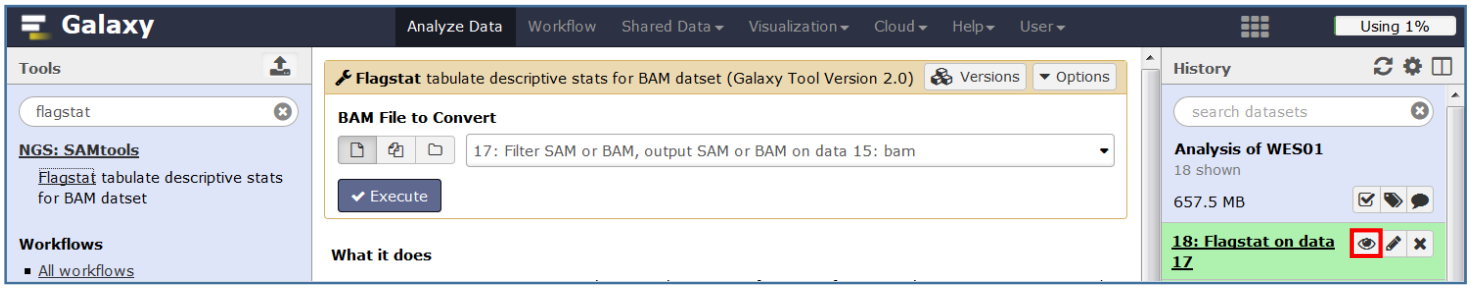




3. In **Tool Pane**: Go to **NGS SAMtools** and rerun Flagstat on the filtered BAM file to generate alignment stats.



4. Click view data to look at the alignment stats for the filtered BAM file.



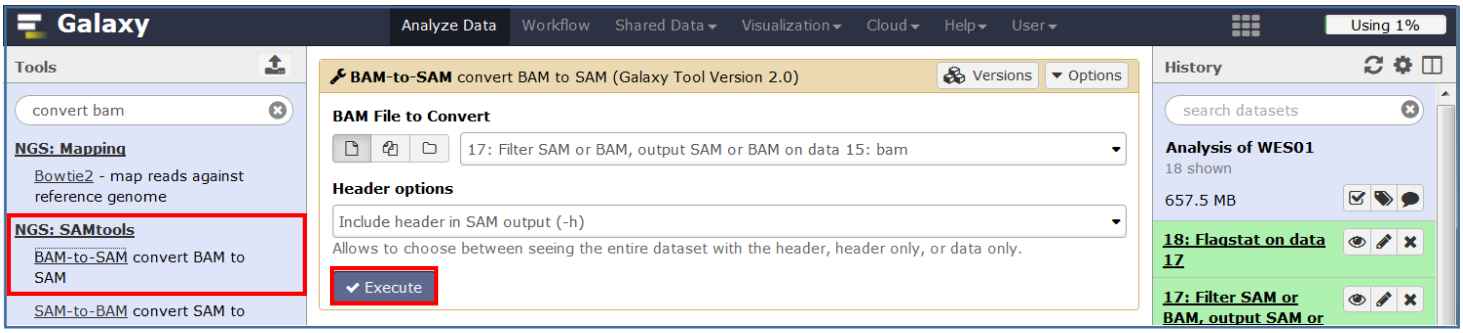
Q7. Use the Flagstat outputs to calculate the number of reads that were filtered out (difference in total read count between the raw and filtered BAM files).

Converting BAM to SAM file

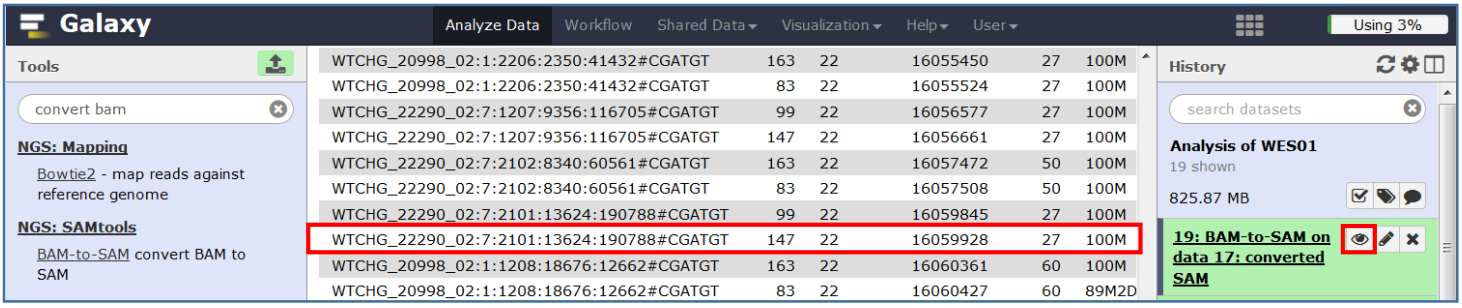
As mentioned above, the BAM file is held in binary format and so it can not be viewed directly. However, the BAM file can be converted to a viewable text format known as a Sequence Alignment/Map file or SAM file. The SAM file format was described in lecture 7 and more details can be found here (http://samtools.github.io/hts-specs/SAMv1.pdf).

Use BAM to SAM to convert the filtered BAM file to a SAM file.

1. In **Tool Pane**: Go to **NGS: SAMtools** > BAM-to-SAM convert BAM to SAM



2. Click view data to view the SAM file



In the SAM file, lines starting with "@" are headers, and those without "@" (as shown above) are read sequences aligned to the reference. The data for the highlighted read (Table 1) show:

- mapped to chromosome 22 at 16,059,928bp
- maps with a quality of phred=27
- Cigar (100M) shows all 100 bases are either matches or mismatches to the reference
- The paired read or mate is mapped to the same chromosome at 16,059,845bp
- The last field (MD:Z:45C54) describes the location of mismatches with respect to the reference. For example, in this read the first 45 bases match the reference followed by a mismatched C and then 54 matches.

Table 1. SAM file format

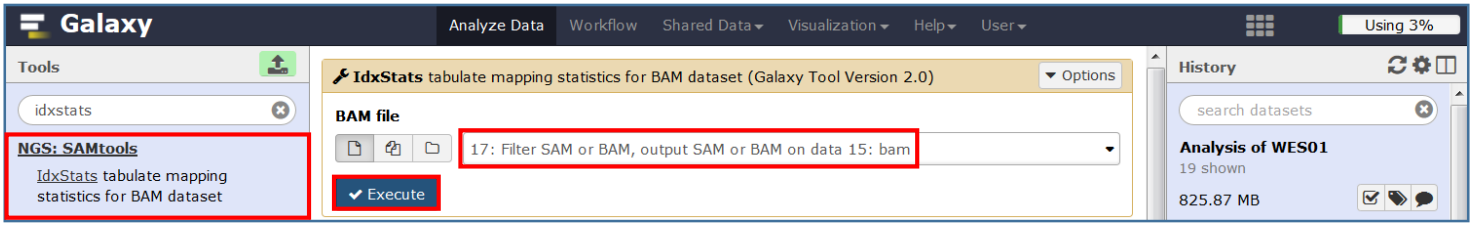
Read name	Flag	Chr.	bp	Mapping quality	Cigar	Mate reference map	Mate position	Template length	Sequence	Quality	
WTCHG_22290_02:7:2101:13624:190788#CGATGT	147	22	16059928	27	100M	=	16059845	-183	CCCACTTTTTAACTTTTGCCAAGAGTAAAGGCTCCCTGGGGCCTCACCAGAAGCTGAATGGATGATAGCACCATGCTTGTACAGCTTGCATTAAAGCTTG	>>>?>@@@@@@@???????>?@?>@??@?@>>??>>>?@@?>@>???@?@@>??>@@?>?????????>@>>@?>>??>??>?>>>??>>??????>>>>=	NM:i:1 MD:Z:45C54 AS:i:95 XS:i:90 XA:Z:14,+19782979,100M,2;

Q8. Use this website (http://broadinstitute.github.io/picard/explain-flags.html) to decode the flag (147) of the read in table 1 and determine which strand of the reference genome it is aligned with.

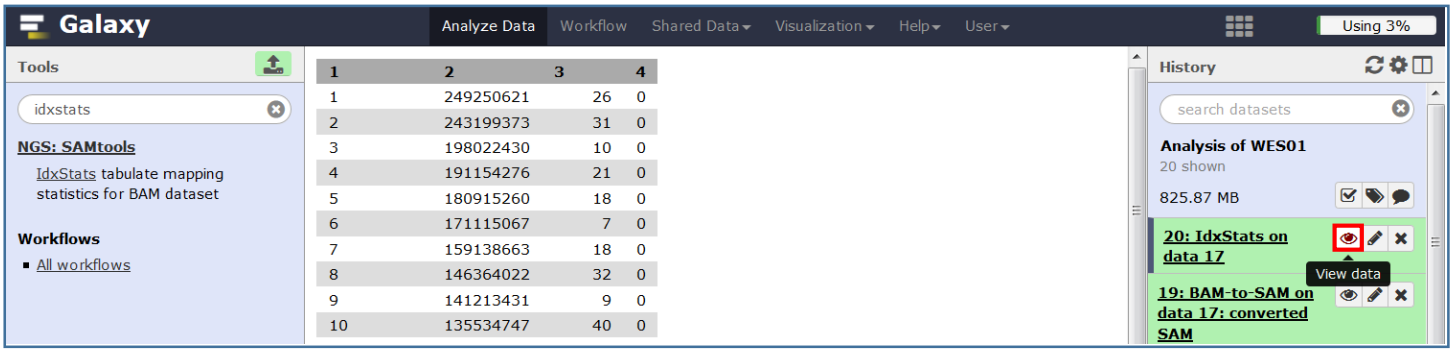
Generate alignment statistics per chromosome

When viewing the SAM file you may have noticed that not all the reads are mapped to chromosome 22 which is surprising as the sequence was initially filtered for reads mapping to chromosome 22 only. Use the IdxStats to generate a breakdown of the number of reads that map to each chromosome or contig.

1. In **Tool Pane**: Go to **NGS: SAMtools** > IdxStats
Select the filtered BAM file (Not the newly created SAM file) and click execute



2. Click view data**.** This will produce a file with 4 columns; 1) Reference sequence identifier. 2) Reference sequence length. 3) Number of mapped reads. 4) Number of placed but unmapped reads (typically unmapped partners of mapped reads)

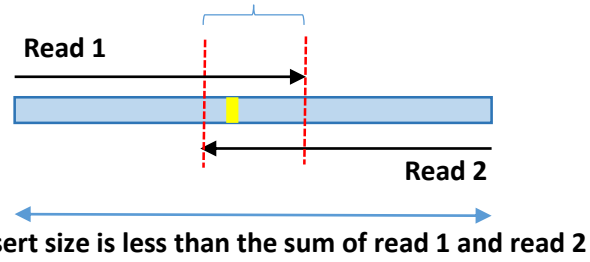


Q9. Use the IdxStats output to calculate the percentage of reads mapping to chromosome 22.

Determine the distribution of insert sizes

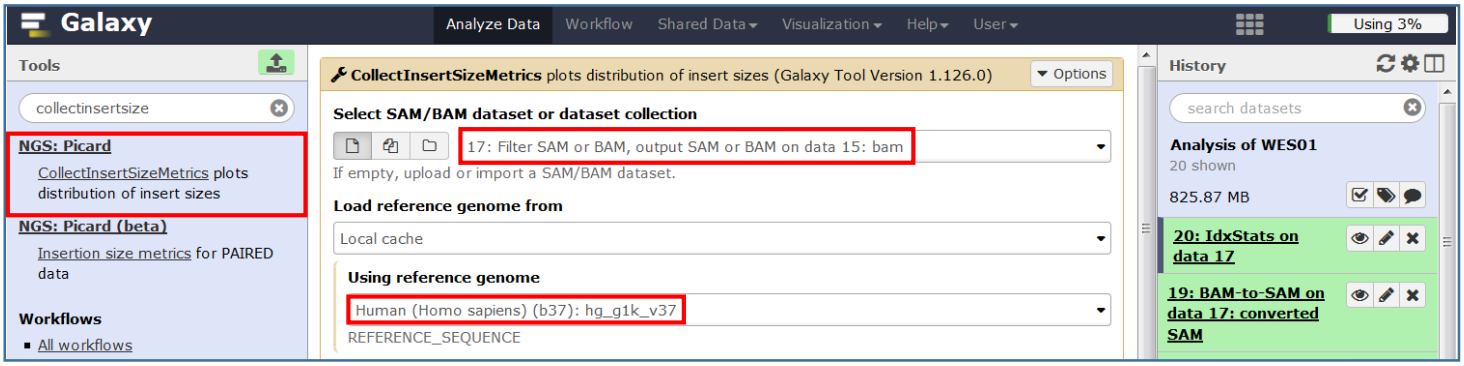
Insert sizes (the region between the 5' ends of the paired reads see Figure 1) are important for correct alignment and variant calling. During alignment with BWA-MEM, we estimated that the mean insert size was 250bp so the program expected reads to be separated by this distance plus or minus the standard deviation. For variant calling, reads are assumed to be independent. However, if an insert is smaller than the sum of the read pairs the reads will overlap and not be independent in the overlapping region (Figure 5). If a PCR error occurs in this overlapping region it will be present in both reads which may result in there being enough evidence to call a variant at this site that is not real.

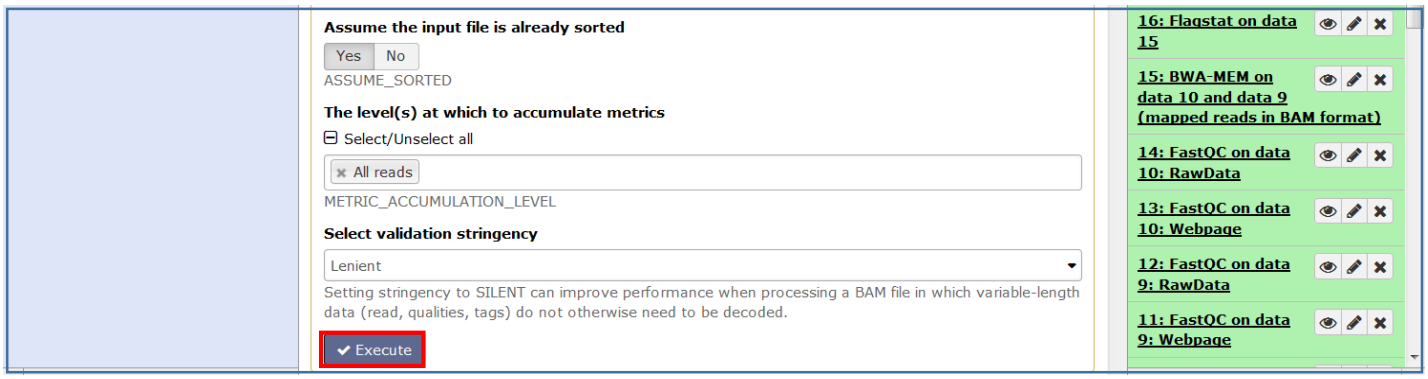
Figure 5. Overalpping reads duplicating a PCR error shown in yellow



Use Picard "CollectInsertSizeMetrics" to produce some statistics and a histogram of the insert size.

1. In **Tool Pane**: Go to **NGS: Picard** > CollectInsertSizeMetrics
Select the filtered bam file, Human hg19 reference genome, don't change the other settings and click execute.

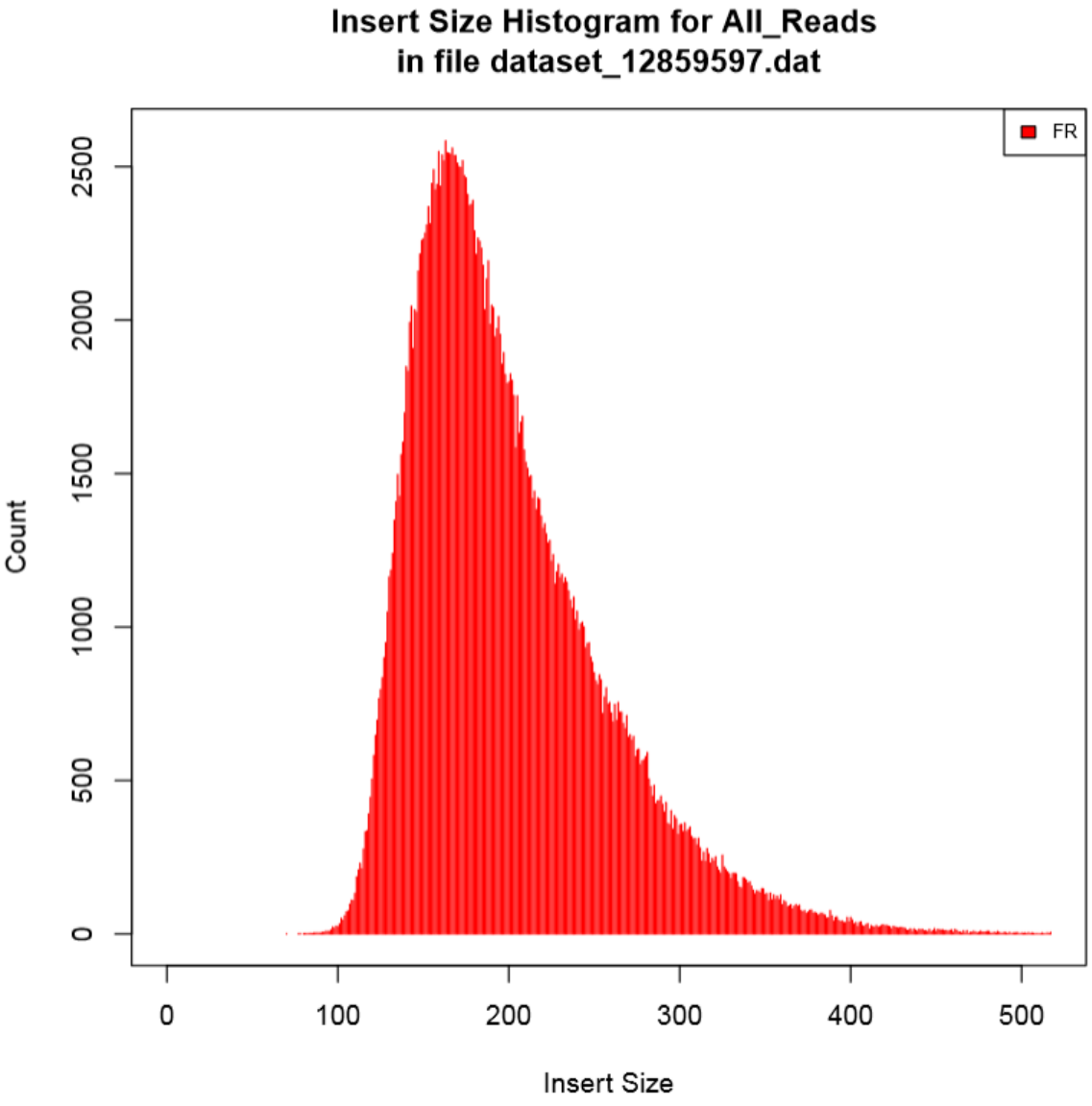




This tool will produce two output files. The first is a tabular output with some statistics and a list of insert sizes and counts.

Q10. View the tabular output and record the mean insert size and standard deviation

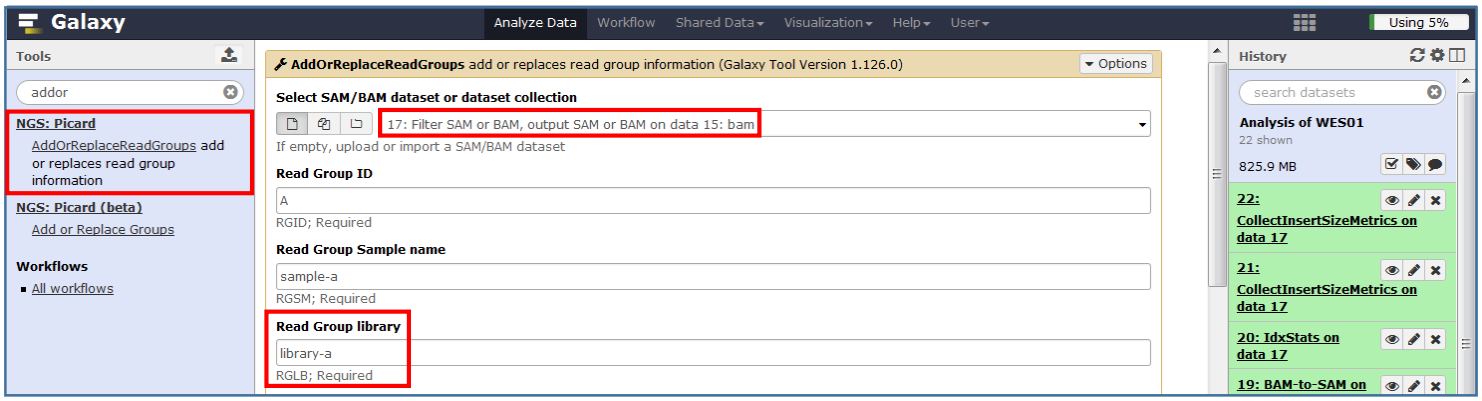
The second output is a .pdf with a insert size histogram, that should look like this;



Add read group information to the BAM file

To use GATK programmes we need to add some read group information to the bam file.

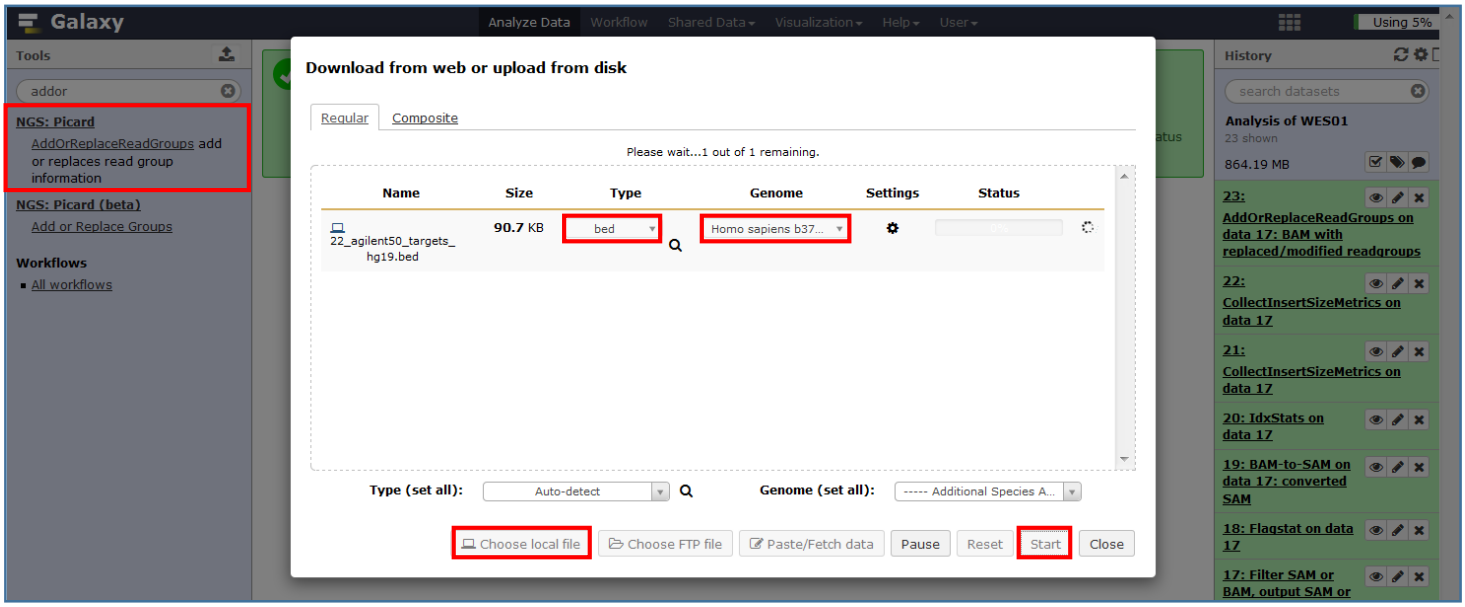
1. In **Tool Pane**: Go to **NGS: Picard** > AddOrReplaceReadGroups
Change read group library to library-a, keep the other settings and click execute



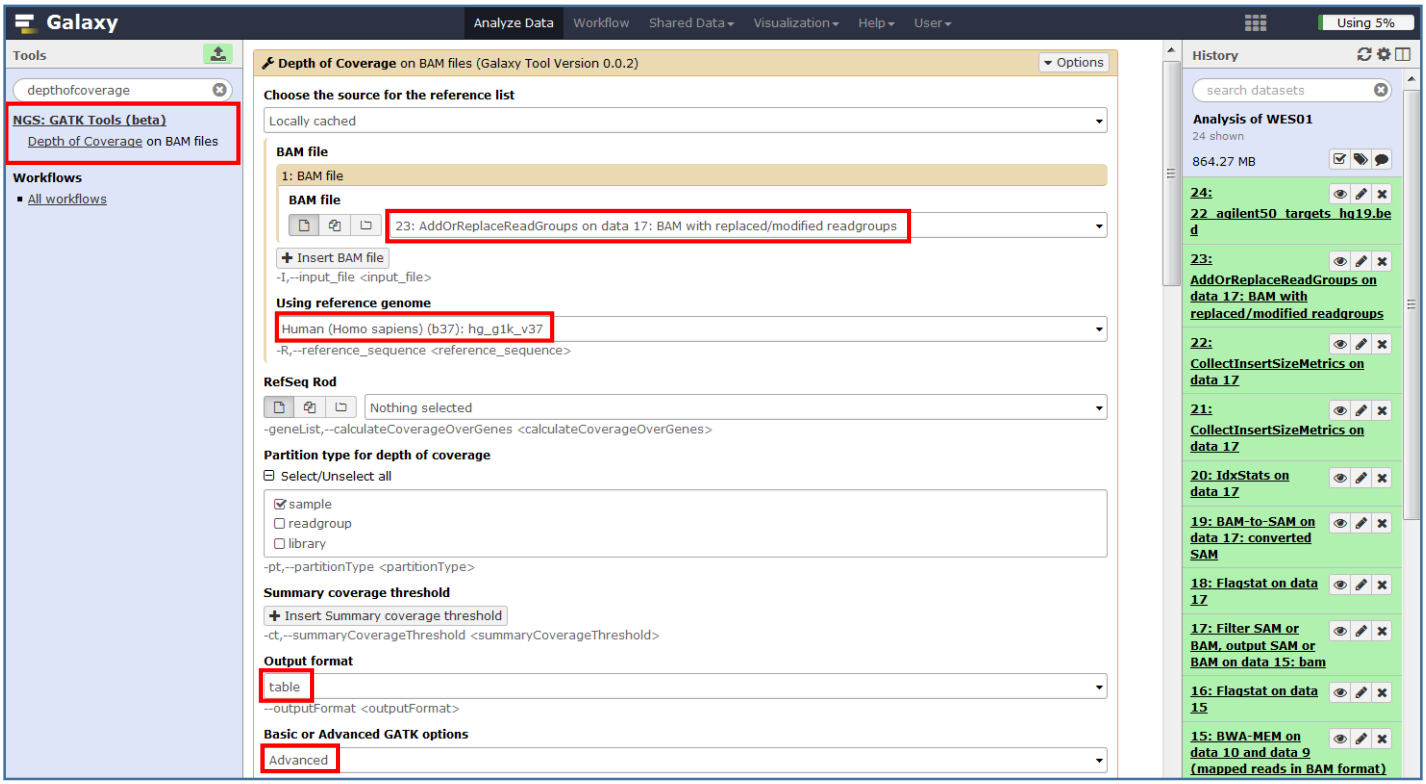
Calculate depth and breadth of coverage

Identification and accuracy of variant calling relies on the depth and breadth of sequence coverage. To calculate coverage, a file in bed format describing the target region is required. Bed files use 3 tab delimited columns of data to describe the target: chromosome, left location (bp), right location (bp). Upload the bed file which describes the exome sequence that has been targeted in your trial data "22_agilent50_targets_hg19.bed".

1. In **Tool Pane**: Go to **Get Data** > Upload File from your computer



2. In **Tool Pane**: Go to **NGS: GATK Tools (beta)** > Depth of Coverage



Q11. What is the mean coverage and percentage of target bases covered by 15 or more reads according to the depth of coverage output?

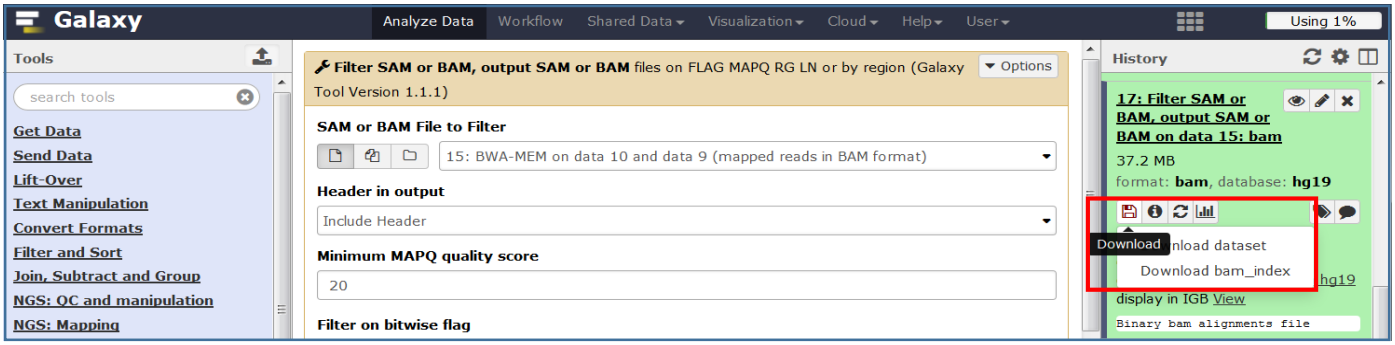
Depth of coverage gives other statistics broken down by regions etc. We're not interested in these so to keep the history clean use the 'X' button to delete the outputs for step 25 and steps 27 to 32.

Visualise alignments using IGV

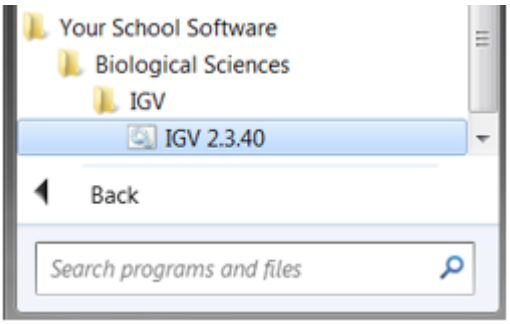
It is useful to look at the aligned data as this is one way to identify variants, assess their quality, and determine their genomic context with respect to other annotated features of the human genome such as genes, repeat sequences, transcription factor binding sites etc. We therefore recommend alignment visualisation before validation of *in-silico* variants by independent sequencing methods. The Integrative Genome Viewer (IGV) is a popular tool for interrogating aligned NGS data (look here for more details on IGV www.ncbi.nlm.nih.gov/pubmed/22517427).

To visualise our data using IGV:

1. Download the filtered BAM dataset and bam_index from step no. 17



2. Launch IGV from Start menu > All Programs > Your School Software > Biological Sciences > IGV > IGV 2.3.40 (This will take some time <5mins as IGV has to load the whole genome, a black window will appear with messages, check this and be patient).



3. When IGV opens, make sure the reference genome is set to hg19 (Figure 6). From the file tab select 'load from file', navigate to the folder with your data, select your bam file and select open. Two new tracks will appear in the left hand pane labelled 'yourfilename.bam coverage' and 'yourfilename.bam'. However, you will not see any aligned reads in the central pane because you are looking at the whole genome, which is too zoomed out and you only have data for the exonic regions of chromosome 22.

4. Load public annotation tracks into IGV. Click 'File' tab and select 'Load from server' in the drop down menu. Select OMIM and dbSNP 1.3.7

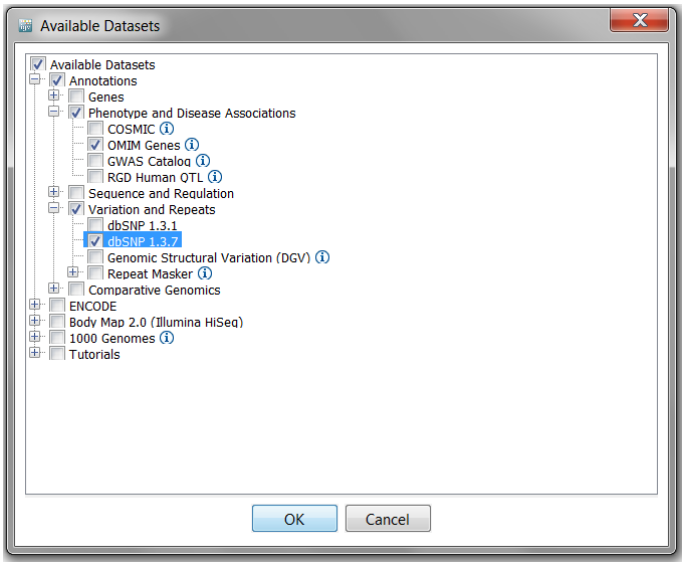
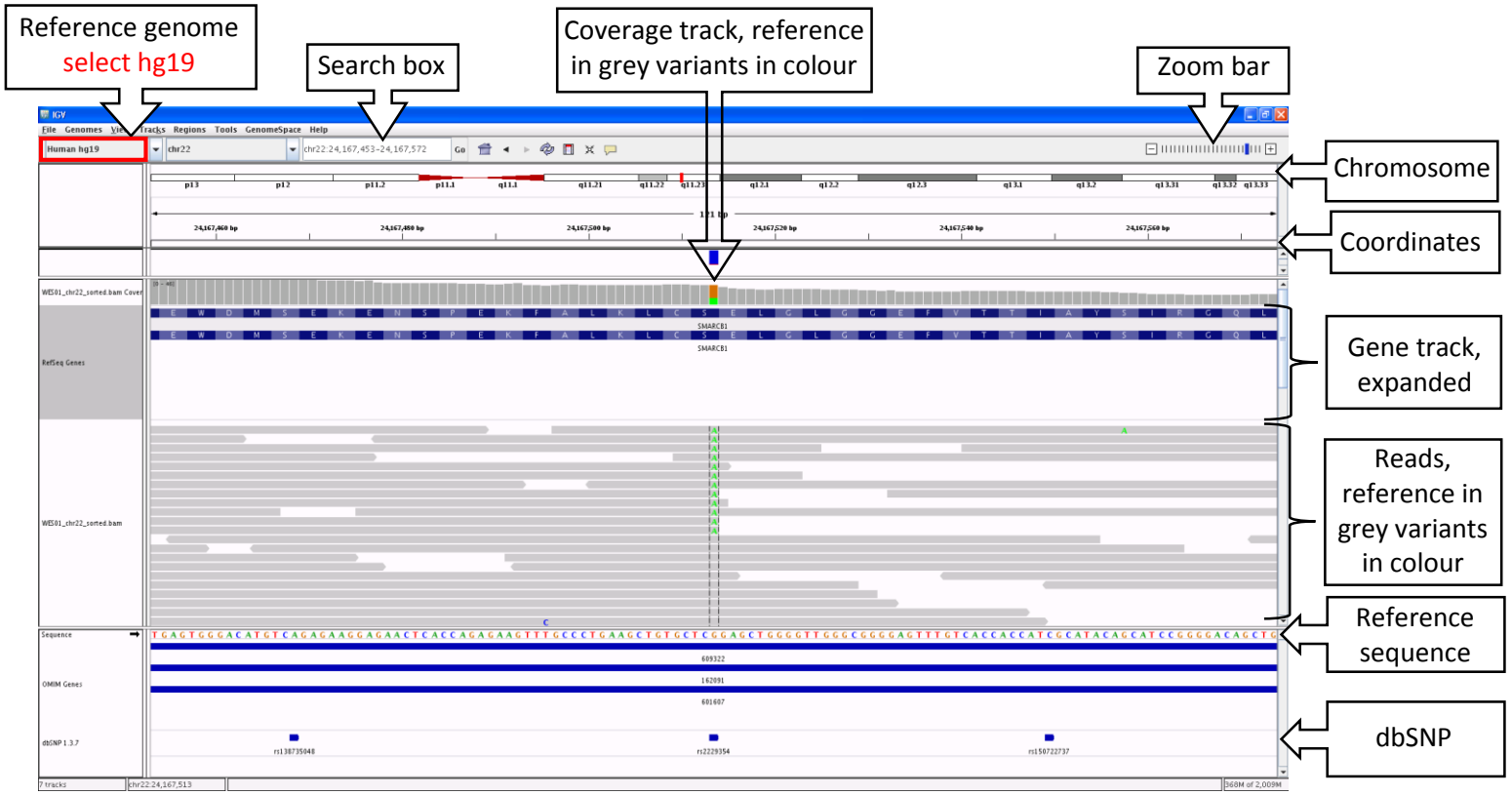


Figure 6. The Integrative Genome Viewer (IGV)

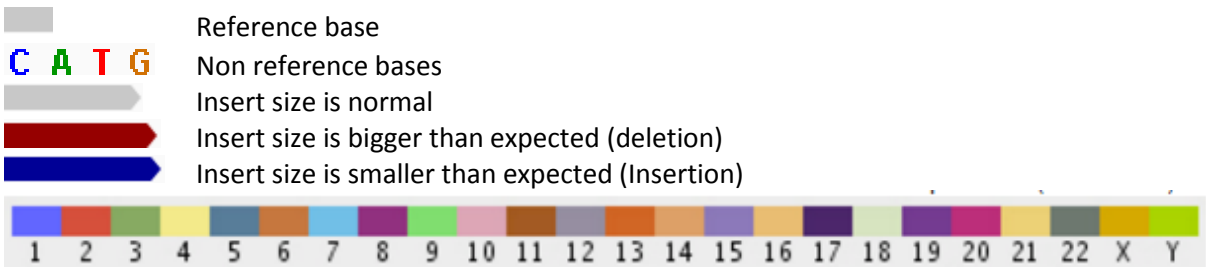


The search box can be used to navigate to features such as a gene or region of interest (Figure 6). SMARCB1 is a tumour suppressor gene that regulates cell cycle, growth and differentiation. An inactivating germline mutation in exon 1 of SMARCB1 has been reported in patients with schwannomatosis (http://omim.org/entry/162091). Schwannomas are mostly benign tumours involving schwann cells that myelinate the axons of nerve cells but can cause problems if the tumour compresses a nerve. There are several cases where people with schwannomatosis have developed hearing loss due to an acoustic neuroma, which is a schwannoma on the vestibular nerve in the brain that is involved in hearing. Mutations in SMARCB1 could, therefore contribute to the patients symptoms.

5. Enter SMARCB1 in the search box and click enter to go to this gene. We can now see data for the whole gene as a coverage profile in the upper track and individual reads below.

Q12: What does the coverage/depth look like in exons and introns and is this expected?

IGV uses a colour code to describe reference sequence, normal reads, mismatched bases and anomalous reads:



For paired end reads that are coded by the chromosome on which their mates can be found.

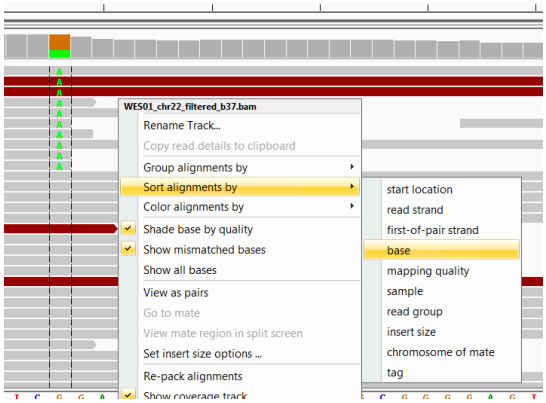
Look here http://www.broadinstitute.org/software/igv/interpreting_insert_size for more details on the colour codes and settings.

If you look closely at the coverage track you will notice some coloured bars which represent SNVs with an alternate allele frequency greater than or equal to 0.2. The allele frequency threshold along with many other settings can be changed by clicking the view tab, selecting preferences from the drop down menu, then click the Alignments tab and changing the Coverage allele-freq threshold.

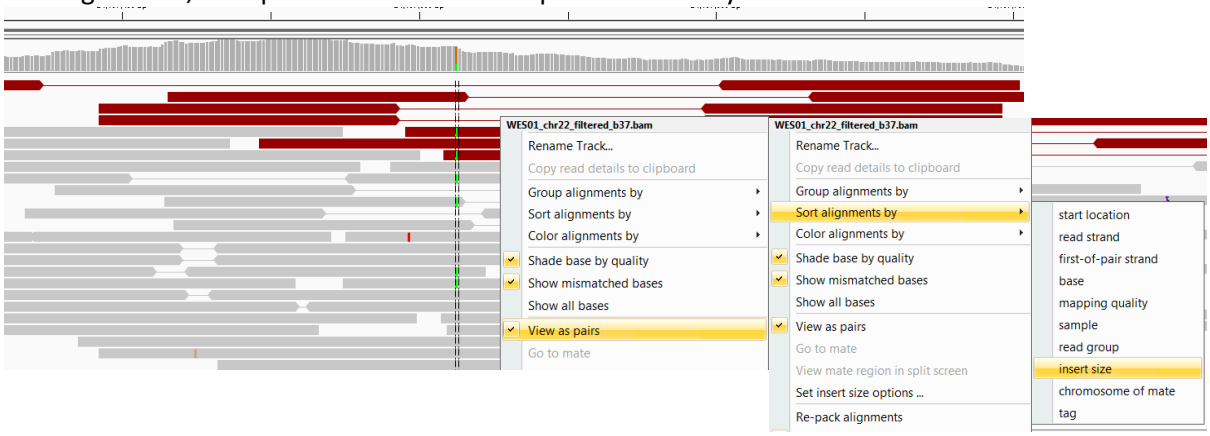
6. Enter the location 'chr22:24,167,513' in the search box to focus on the data for a particular variant. You can now see the coloured bar consists of two colours representing the two alleles and their height corresponds with the allele frequency.

There are many ways to present the data in IGV, which help to explore different aspects of the data. For variants, it helps to sort the alignments by base.

Right click and select sort alignments by base.



For large indels, it helps to view the reads as pairs and sort by insert size.



Q13. Mouse over the coverage track for the SNV at 24,167,513bp and record its alleles, number of reads with the reference allele, no. reads with the alternative allele, gene, amino acid that it occurs in, and the rsid (rs#) from dbSNP if there is one.

Q14. Is the variant at 24,167,513bp likely to contribute to patients symptoms?

Congratulations you finished the exercise!

In the next practical we will investigate automated methods of variant calling.