

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

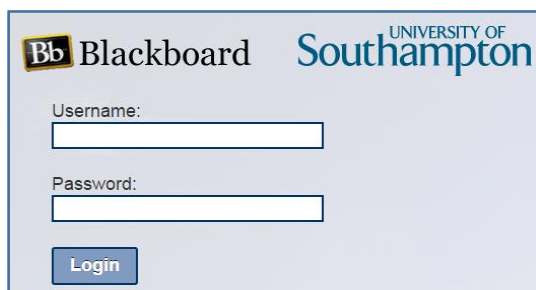
Trial data

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.

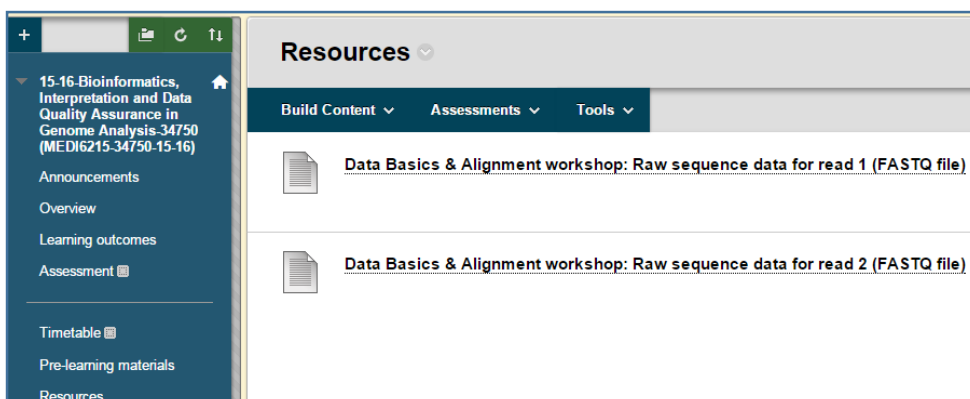
The patients exome was sequenced using the paired-end method on an Illumina HiSeq 2000 following target enrichment by Agilent SureSelect. Your aim over the next three practicals is to analyse this data and determine if there are any disease causing mutations, and if so what disease is implicated. While following the tasks below, think about the genes that should be prioritised for mutation screening given the patients symptoms.

Let's begin

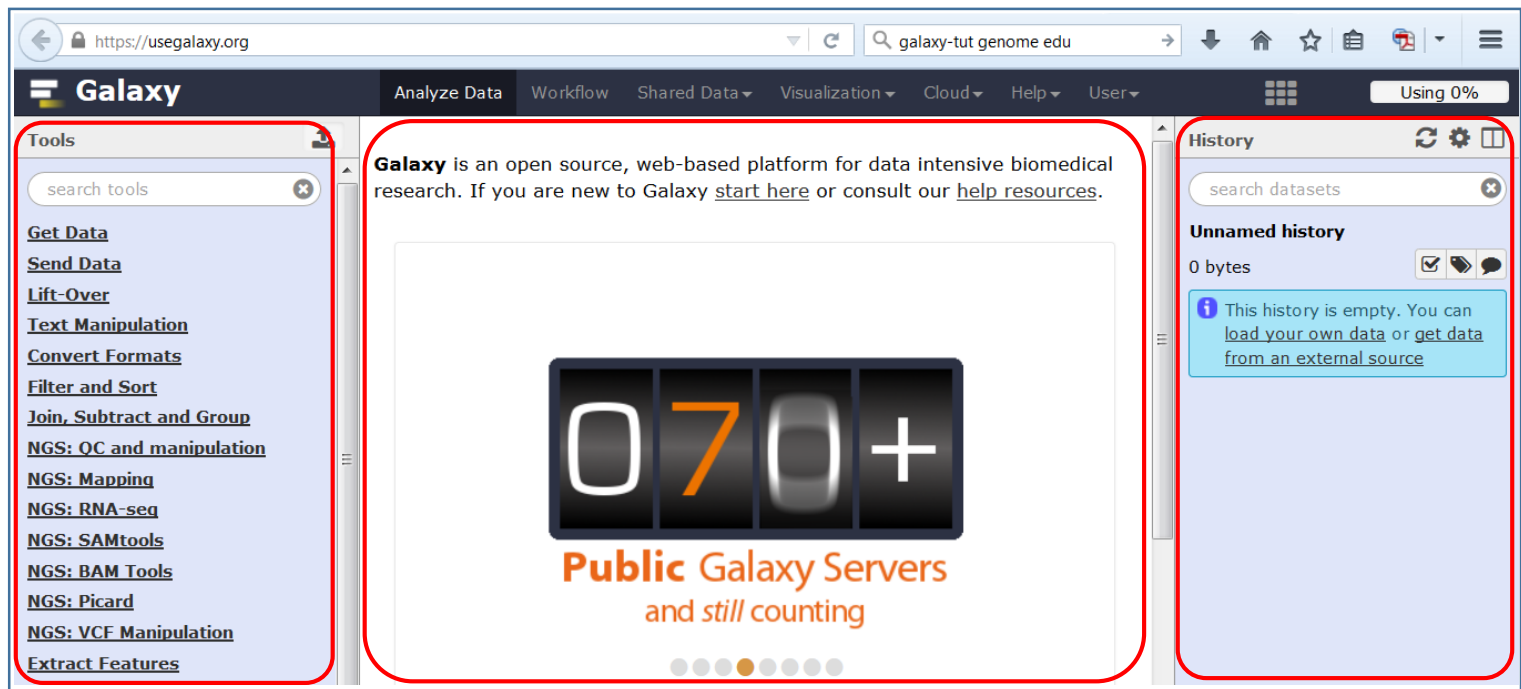
1. Login to blackboard: <https://blackboard.soton.ac.uk/>



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

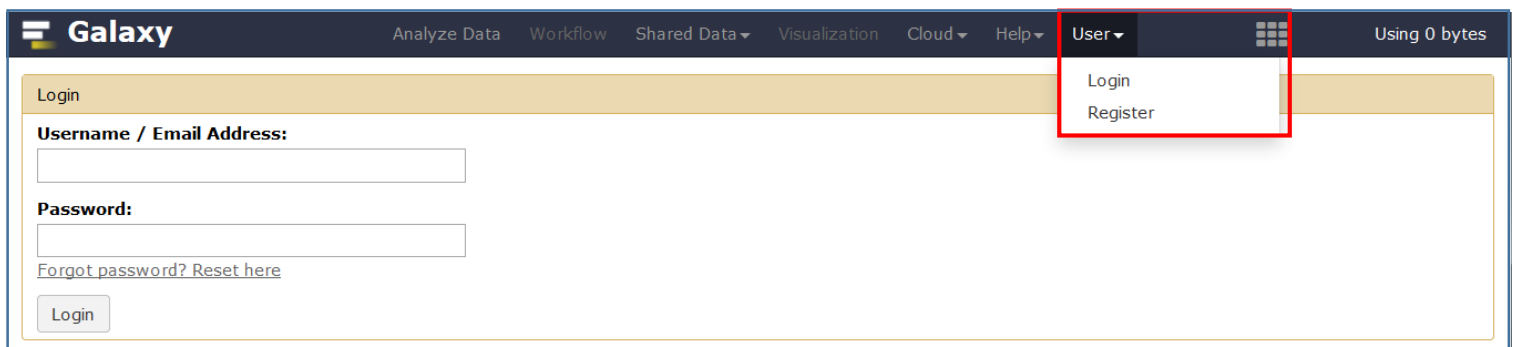


3. Go to <https://usegalaxy.org/>

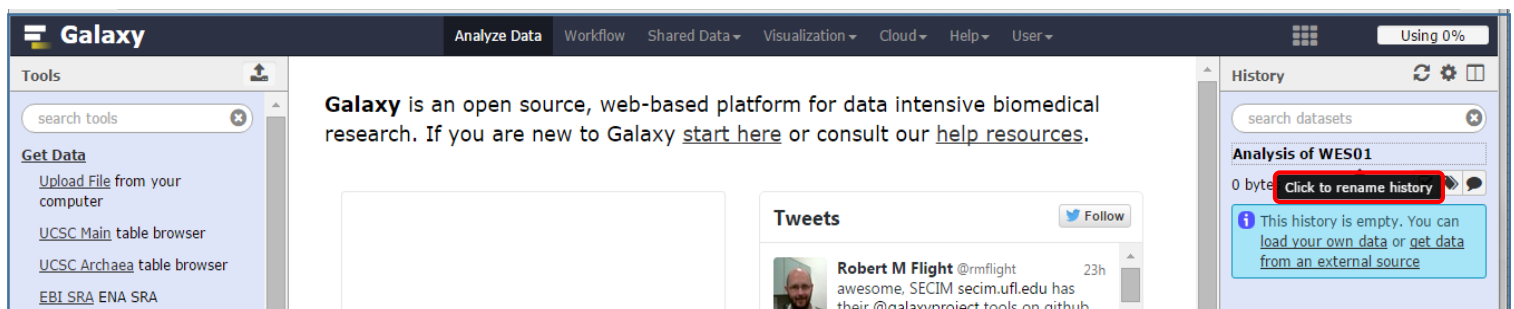


The available “Tools” are in the left tool pane, the central panel is your working area where you can visualise your data, select tool options and execute jobs. Your “History” is shown in the right hand panel. Here you will see your data, results and a history of all the tools that you have used.

4. Register for a galaxy account and login

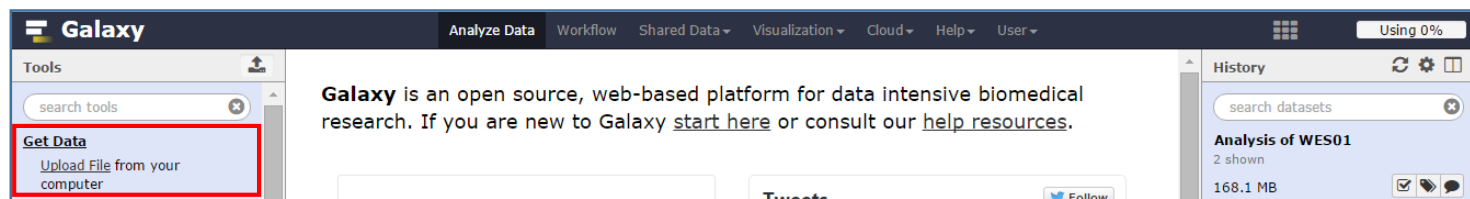


5. Rename the history for this session (Click on name, type new name, press return)

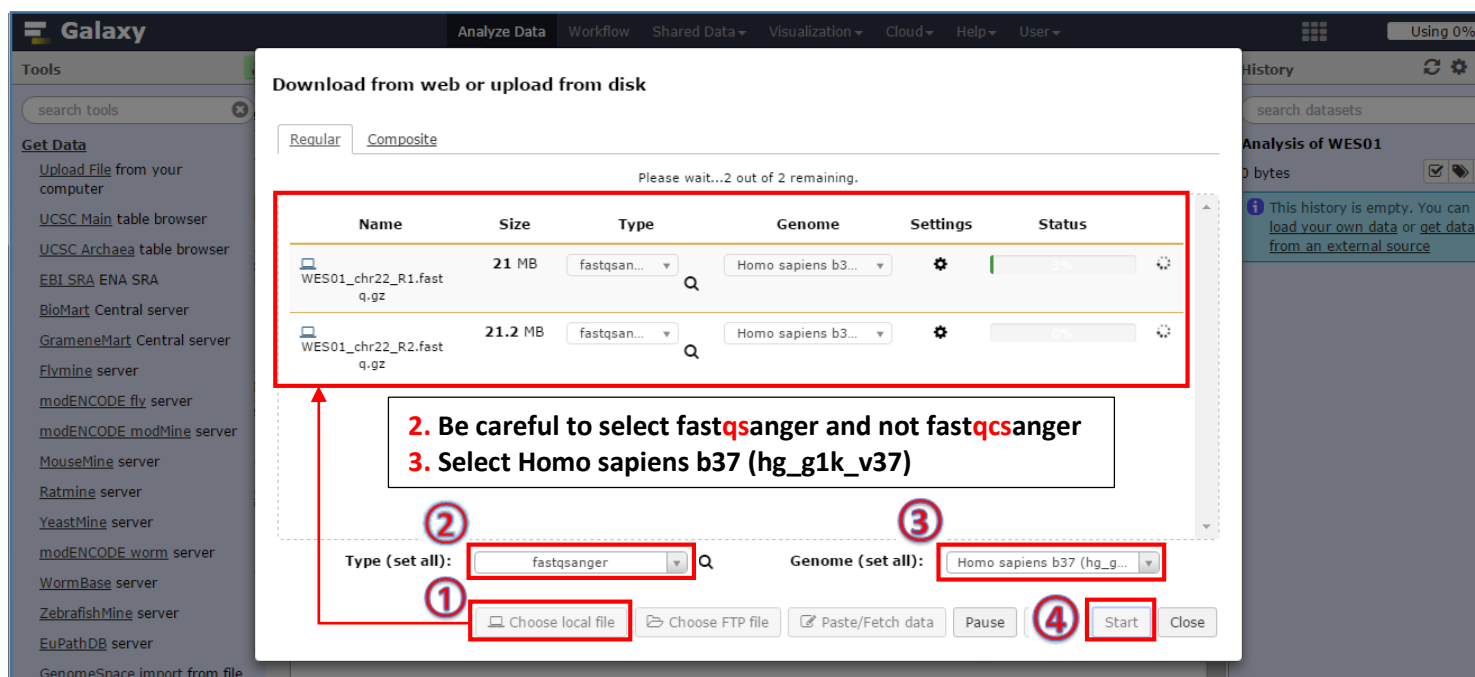


Uploading Files

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

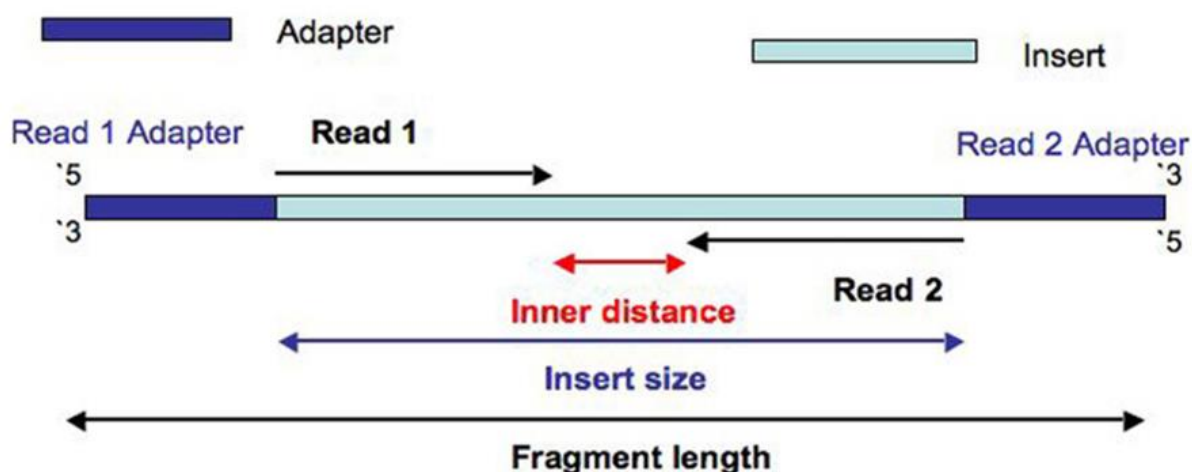


2. Click **Upload File**



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



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

The screenshot shows the Galaxy web interface. In the 'Tools' pane on the left, 'NGS: QC and manipulation' is selected, and 'FastQC Read Quality reports' is highlighted. The main panel displays a dataset preview for 'WES01' with a warning: 'This dataset is large and only the first megabyte is shown below.' The 'History' pane on the right shows the dataset 'WES01' and its analysis, including '2: WES01_chr22_R2.f' and '1: WES01_chr22_R1.f'.

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

The screenshot shows the 'FastQC Read Quality reports' tool interface. The 'Short read data from your current history' section shows two datasets selected: '2: WES01_chr22_R2.fastq' and '1: WES01_chr22_R1.fastq'. A red box highlights the 'Execute' button at the bottom.

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.

The screenshot shows the Galaxy web interface. The 'History' pane on the right shows the 'FastQC' jobs. The jobs are listed in a table with columns for job ID, name, status, and actions. The jobs are: '6: FastQC on data 2: RawData' (green), '5: FastQC on data 2: Webpage' (green), '4: FastQC on data 1: RawData' (grey), '3: FastQC on data 1: Webpage' (grey), '2: WES01_chr22_R2.f' (green), and '1: WES01_chr22_R1.f' (green). The 'Tools' pane on the left shows 'NGS: QC and manipulation' selected.

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

The screenshot shows the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Cloud, Help, and User. The left sidebar lists various NGS QC tools under the category 'NGS: QC and manipulation'. The central panel displays the 'FastQC Report' for 'WES01_chr22_R1.fastq', dated 'Sun 27 Sep 2015'. The report includes a 'Summary' section with a list of checks: Basic Statistics, Per base sequence quality, Per tile sequence quality, Per sequence quality scores, Per base sequence content, Per sequence GC content, and Per base N content. The right sidebar shows a 'History' of analyses, with '3: FastQC on data 1: Webpage' highlighted in red.

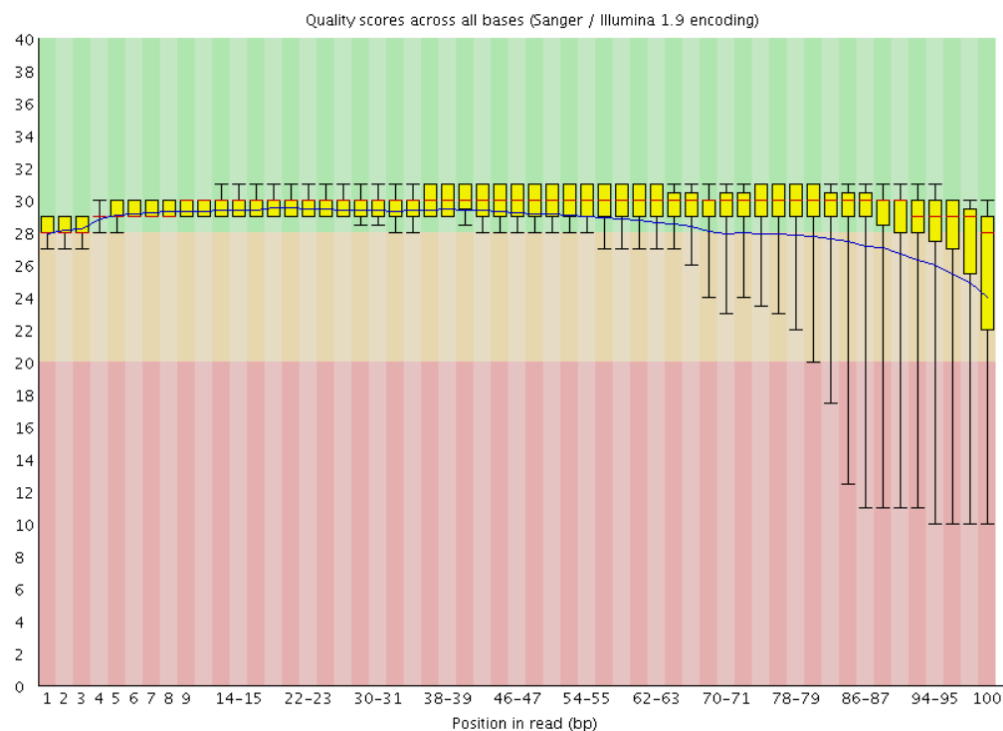
Use the FastQC reports on data 1 and 2 to answer:

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?

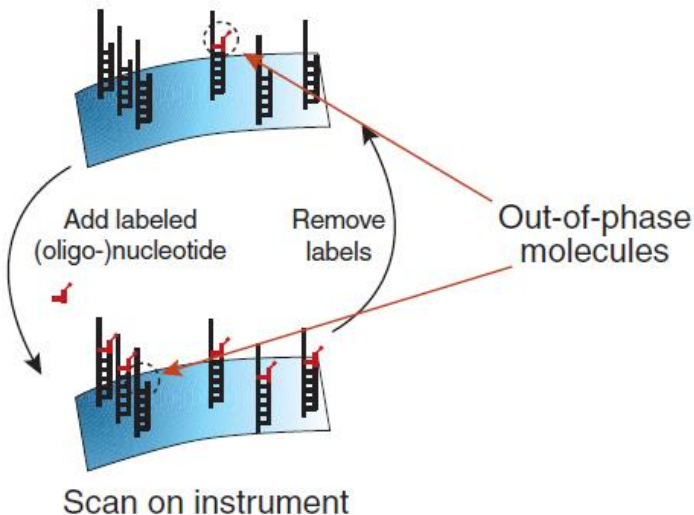
Figure 3. Per base sequence quality for WES01_chr22_R1.fastq



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**

The screenshot shows the Galaxy web interface. The 'Tools' pane on the left lists 'NGS: QC and manipulation' and 'FASTQ joiner on paired end reads'. The 'FASTQ joiner' tool is selected, and its configuration is shown in the main pane. The tool is configured to join two input files: '1: WES01_chr22_R1.fastq' and '2: WES01_chr22_R2.fastq'. A warning box is overlaid on the interface, stating: 'Be careful to select: WES01_chr22_R1.fastq WES01_chr22_R2.fastq'. The interface also shows the 'History' pane on the right, which lists previous jobs, and the 'FASTQ joiner' tool description at the bottom.

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.

The screenshot shows the Galaxy web interface with the 'Filter by quality' tool selected. The tool configuration is as follows:

- Library to filter:** 7: FASTQ joiner on data 2 and data 1
- Quality cut-off value:** 20
- Percent of bases in sequence that must have quality equal to / higher than cut-off value:** 90
- Execute** button is highlighted.

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

The screenshot shows the details of the dataset '8: Filter by quality on data 7'. The dataset size is 121.3 MB. The format is fastqsanger, database: hg19. The quality cut-off is 20, and the minimum percentage is 90.

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**

The screenshot shows the Galaxy web interface with the 'FASTQ splitter' tool selected. The tool configuration is as follows:

- FASTQ reads:** 8: Filter by quality on data 7
- Execute** button is highlighted.

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

The two screenshots show the results of the FASTQ splitter tool. Both show 67.5 MB files, format: fastqsanger, database: hg_g1k_v37, and a message: 'Split 285237 of 285237 reads (100.00%).'

BWA-MEM - map medium and long reads (> 100 bp) against reference genome

Map with BWA for Illumina

NGS: SAMtools

Stats generate statistics for BAM dataset

Filter SAM on bitwise flag values

NGS: Picard

FilterSamReads include or exclude aligned and unaligned reads and read lists

Workflows

- All workflows

Enter mean, standard deviation, max, and min for insert lengths.

200

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?

Set

Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple datasets. See help below for more details

Specify readgroup ID

readgroup1

This value must be unique among multiple samples in your experiment

Specify readgroup sample name (SM)

blood

This value should be descriptive

Select analysis mode

1.Simple Illumina mode

Job Resource Parameters

Use default job resource parameters

Execute

13: FastQC on data 10: Webpage

12: FastQC on data 9: RawData

11: FastQC on data 9: Webpage

10: FASTQ splitter on data 8

9: FASTQ splitter on data 8

8: Filter by quality on data 7

7: FASTQ joiner on data 2 and data 1

6: FastQC on data 2: RawData

5: FastQC on data 2: Webpage

4: FastQC on data 1: RawData

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.

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

BWA

NGS: QC and manipulation

Convert SOLiD output to fastq

NGS: Mapping

BWA - map short reads (< 100 bp) against reference genome

BWA-MEM - map medium and

1 job has been successfully added to the queue - resulting in the following datasets:

15: BWA-MEM on data 10 and data 9 (mapped reads in BAM format)

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

search datasets

Analysis of WES01

15 shown

620.2 MB

15: BWA-MEM on data 10 and data 9 (mapped reads in BAM format)

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

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

flagstat

NGS: SAMtools

Flagstat tabulate descriptive stats for BAM dataset

Flagstat tabulate descriptive stats for BAM dataset (Galaxy Tool Version 2.0)

BAM File to Convert

15: BWA-MEM on data 10 and data 9 (mapped reads in BAM format)

Execute

History

search datasets

Analysis of WES01

15 shown

620.2 MB

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

The screenshot shows the Galaxy web interface. A green notification box states: "1 job has been successfully added to the queue - resulting in the following datasets: 16: Flagstat on data 15". Below this, it says: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." On the right, the History pane shows a list of datasets, including "16: Flagstat on data 15" which is highlighted with a red box around the 'view data' icon.

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

The screenshot shows the Galaxy web interface with the 'Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region (Galaxy)' tool selected. The tool version is 1.1.1. The configuration is as follows:

- SAM or BAM File to Filter:** 15: BWA-MEM on data 10 and data 9 (mapped reads in BAM format)
- Header in output:** Include Header
- Minimum MAPQ quality score:** 20
- Filter on bitwise flag:** yes
- Only output alignments with all of these flag bits set:**
 - ☐ Read is paired
 - ☐ Read is mapped in a proper pair
 - ☒ The read is unmapped
 - ☐ The mate is unmapped
 - ☐ Read strand
 - ☐ Mate strand
 - ☐ Read is the first in a pair
 - ☐ Read is the second in a pair
 - ☒ The alignment or this read is not primary
 - ☒ The read fails platform/vendor quality checks
 - ☒ The read is a PCR or optical duplicate

A red box highlights the 'Skip alignments with any of these flag bits set' section. A warning icon (yellow triangle with an exclamation mark) is overlaid on the right side of the interface, with the text: "Be careful to select options in the 'Skip alignments' section".

Select the output format

bam

✓ Execute

What it does

This tool uses the samtools view command in [SAMTools](#) toolkit to filter a SAM or BAM file on the MAPQ (mapping quality), FLAG bits, Read Group, Library, or region.

History

- 9: FASTQ splitter on data 8
- 8: Filter by quality on data 7
- 7: FASTQ joiner on data 2 and data 1
- 6: FastQC on data 2: RawData

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

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

flagstat

NGS: SAMtools

Flagstat tabulate descriptive stats for BAM dataset

Flagstat tabulate descriptive stats for BAM dataset (Galaxy Tool Version 2.0)

BAM File to Convert

17: Filter SAM or BAM, output SAM or BAM on data 15: bam

✓ Execute

History

search datasets

Analysis of WES01

17 shown

657.5 MB

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

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

flagstat

NGS: SAMtools

Flagstat tabulate descriptive stats for BAM dataset

Workflows

All workflows

Flagstat tabulate descriptive stats for BAM dataset (Galaxy Tool Version 2.0)

BAM File to Convert

17: Filter SAM or BAM, output SAM or BAM on data 15: bam

✓ Execute

History

search datasets

Analysis of WES01

18 shown

657.5 MB

18: Flagstat on data 17

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

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

convert bam

NGS: Mapping

Bowtie2 - map reads against reference genome

NGS: SAMtools

BAM-to-SAM convert BAM to SAM

SAM-to-BAM convert SAM to

BAM-to-SAM convert BAM to SAM (Galaxy Tool Version 2.0)

BAM File to Convert

17: Filter SAM or BAM, output SAM or BAM on data 15: bam

Header options

Include header in SAM output (-h)

Allows to choose between seeing the entire dataset with the header, header only, or data only.

✓ Execute

History

search datasets

Analysis of WES01

18 shown

657.5 MB

18: Flagstat on data 17

17: Filter SAM or BAM, output SAM or BAM on data 15: bam

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)

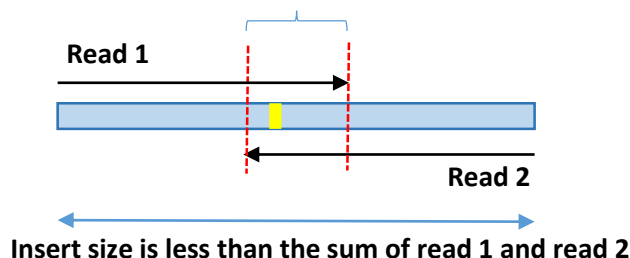
Galaxy				
Analyze Data Workflow Shared Data Visualization Help User				
Tools				
idxstats				
NGS: SAMtools				
IdxStats tabulate mapping statistics for BAM dataset				
Workflows				
All workflows				
1	2	3	4	
1	249250621	26	0	
2	243199373	31	0	
3	198022430	10	0	
4	191154276	21	0	
5	180915260	18	0	
6	171115067	7	0	
7	159138663	18	0	
8	146364022	32	0	
9	141213431	9	0	
10	135534747	40	0	

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. Overlapping 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.

Galaxy				
Analyze Data Workflow Shared Data Visualization Help User				
Tools				
collectinsize				
NGS: Picard				
CollectInsertSizeMetrics plots distribution of insert sizes				
NGS: Picard (beta)				
Insertion size metrics for PAIRED data				
Workflows				
All workflows				

CollectInsertSizeMetrics plots distribution of insert sizes (Galaxy Tool Version 1.126.0)				
Options				
Select SAM/BAM dataset or dataset collection				
17: Filter SAM or BAM, output SAM or BAM on data 15: bam				
If empty, upload or import a SAM/BAM dataset.				
Load reference genome from				
Local cache				
Using reference genome				
Human (Homo sapiens) (b37): hg_g1k_v37				
REFERENCE_SEQUENCE				

Assume the input file is already sorted

ASSUME_SORTED

The level(s) at which to accumulate metrics

☐ Select/Unselect all

All reads

METRIC_ACCUMULATION_LEVEL

Select validation stringency

Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded.

16: Flagstat on data 15

15: BWA-MEM on data 10 and data 9 (mapped reads in BAM format)

14: FastQC on data 10: RawData

13: FastQC on data 10: Webpage

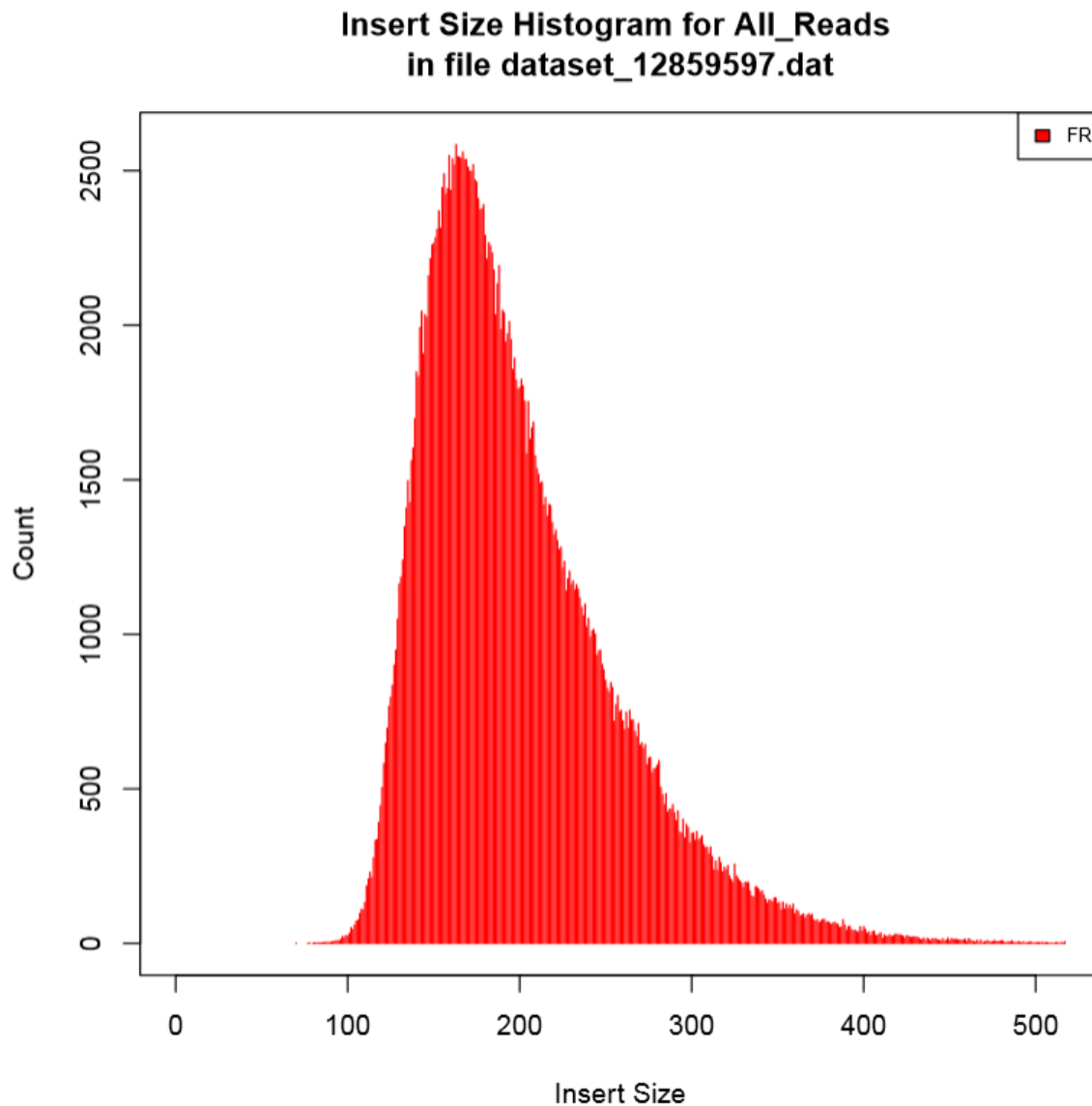
12: FastQC on data 9: RawData

11: FastQC on data 9: Webpage

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

The screenshot shows the Galaxy web interface. On the left, the 'Tools' sidebar has 'NGS: Picard' selected, with 'AddOrReplaceReadGroups add or replaces read group information' highlighted. The main panel displays the 'AddOrReplaceReadGroups' tool configuration. The 'Select SAM/BAM dataset or dataset collection' dropdown is set to '17: Filter SAM or BAM, output SAM or BAM on data 15: bam'. The 'Read Group ID' is 'A', 'Read Group Sample name' is 'sample-a', and 'Read Group library' is 'library-a'. The 'Select validation stringency' dropdown is set to 'Lenient'. The 'Execute' button is highlighted in red.

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

The screenshot shows the Galaxy web interface with a 'Download from web or upload from disk' dialog box open. The 'Regular' tab is selected. The dialog shows a table with the following data:

Name	Size	Type	Genome	Settings	Status
22_agilent50_targets_hg19.bed	90.7 KB	bed	Homo sapiens b37...		

The 'Choose local file' button is highlighted in red. The 'Start' button is also highlighted in red.

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

The screenshot displays the Galaxy web interface for the 'Depth of Coverage' tool. The tool is configured with the following parameters:

- Tool Name:** Depth of Coverage on BAM files (Galaxy Tool Version 0.0.2)
- Choose the source for the reference list:** Locally cached
- BAM file:** 1: BAM file; 23: AddOrReplaceReadGroups on data 17: BAM with replaced/modified readgroups
- Using reference genome:** Human (Homo sapiens) (b37): hg_g1k_v37
- RefSeq Rod:** Nothing selected
- Partition type for depth of coverage:** Select/Unselect all; sample
- Summary coverage threshold:** table
- Output format:** table
- Basic or Advanced GATK options:** Advanced

The 'Execute' button is highlighted in red. The right sidebar shows a history of datasets, including 'Analysis of WES01' and various intermediate files.

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

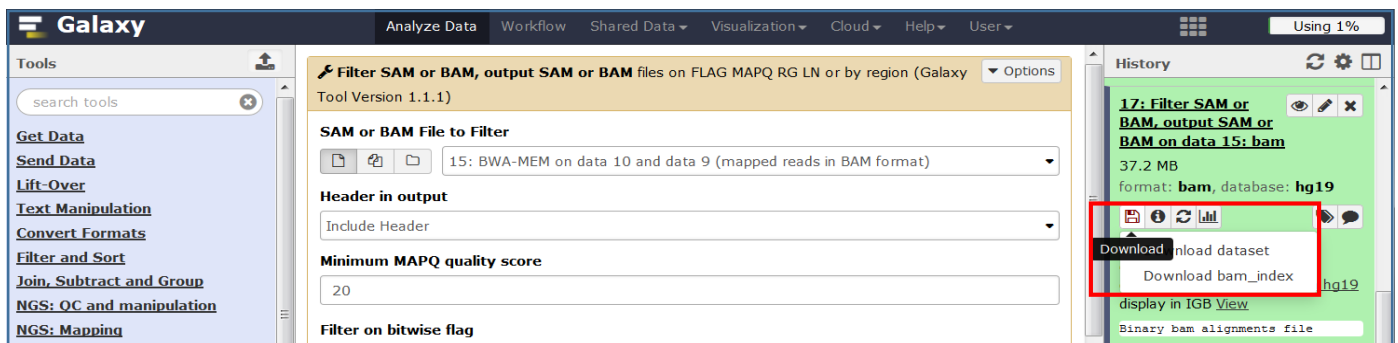
Depth of coverage gives other statistics broken down by regions etc. We're not interested in these so to keep the history clean use the 'X' button to delete the outputs for step 25 and steps 27 to 32.

Visualise alignments using IGV

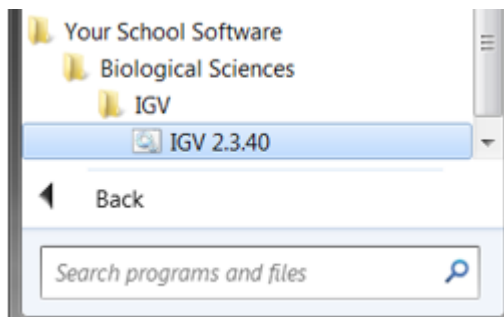
It is useful to look at the aligned data as this is one way to identify variants, assess their quality, and determine their genomic context with respect to other annotated features of the human genome such as genes, repeat sequences, transcription factor binding sites etc. We therefore recommend alignment 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

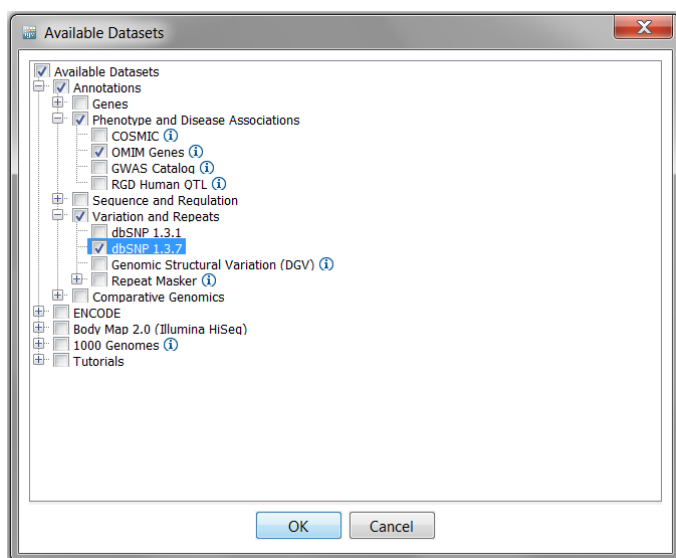
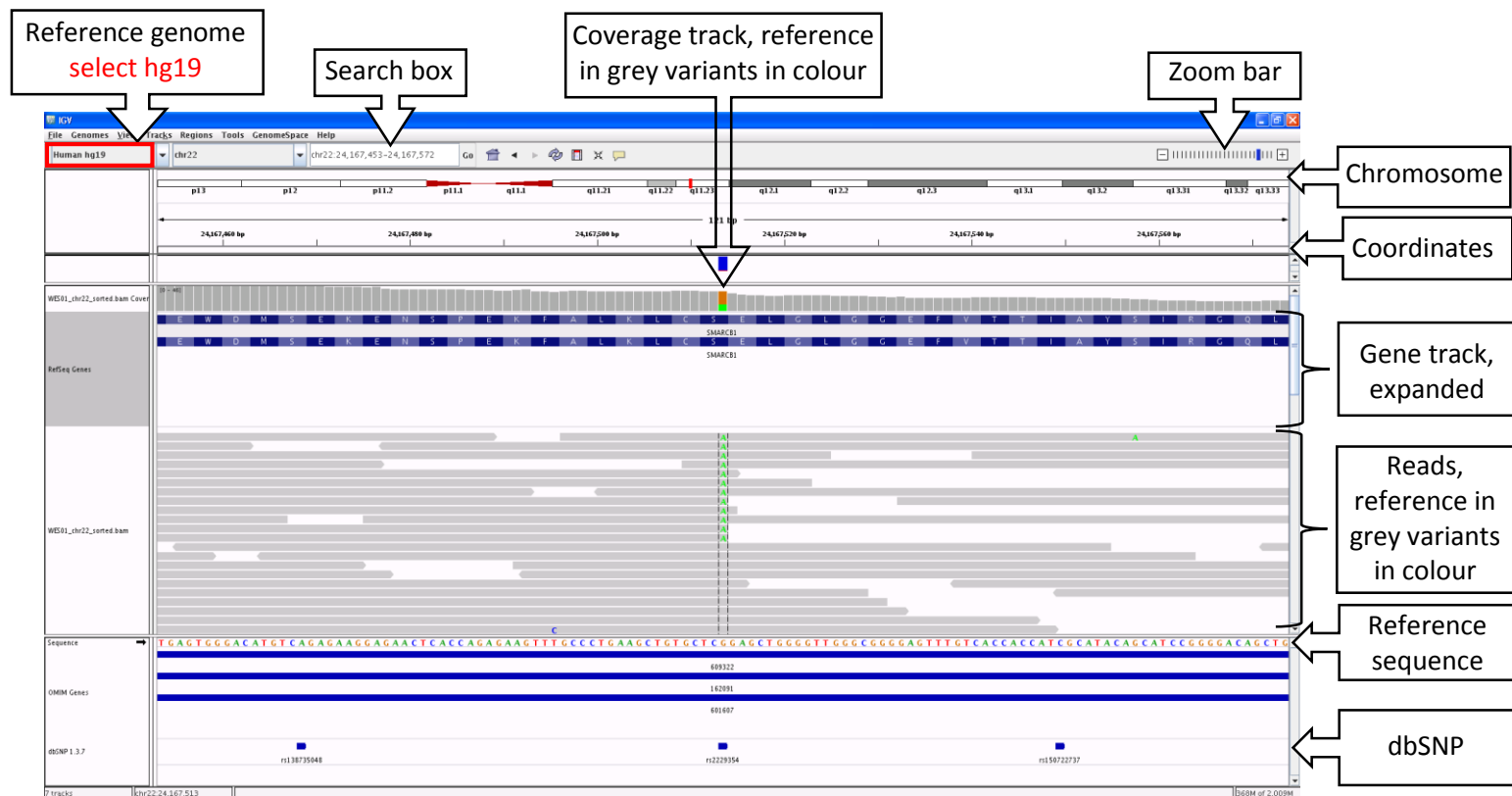


Figure 6. The Integrative Genome Viewer (IGV)

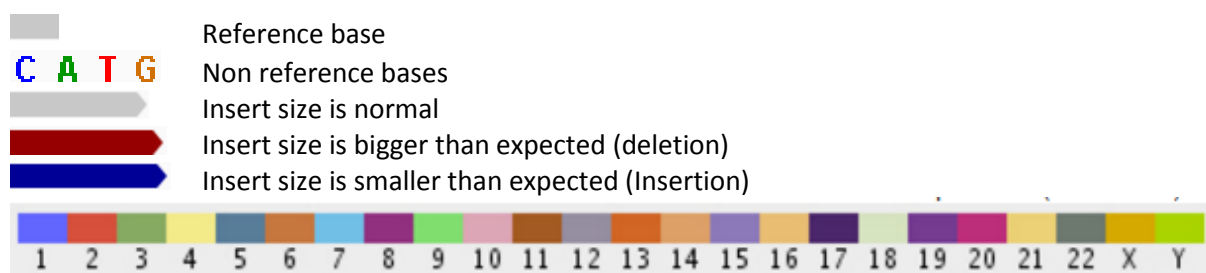


The search box can be used to navigate to features such as a gene or region of interest (Figure 6). SMARCB1 is a tumour suppressor gene that regulates cell cycle, growth and differentiation. An inactivating germline mutation in exon 1 of SMARCB1 has been reported in patients with schwannomatosis (<http://omim.org/entry/162091>). 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.

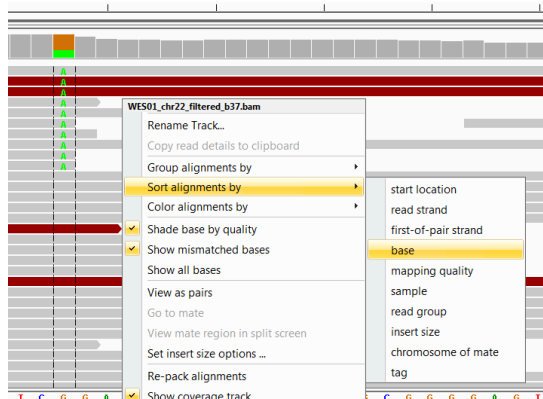
Look here http://www.broadinstitute.org/software/igv/interpreting_insert_size for more details on the colour codes and settings.

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.

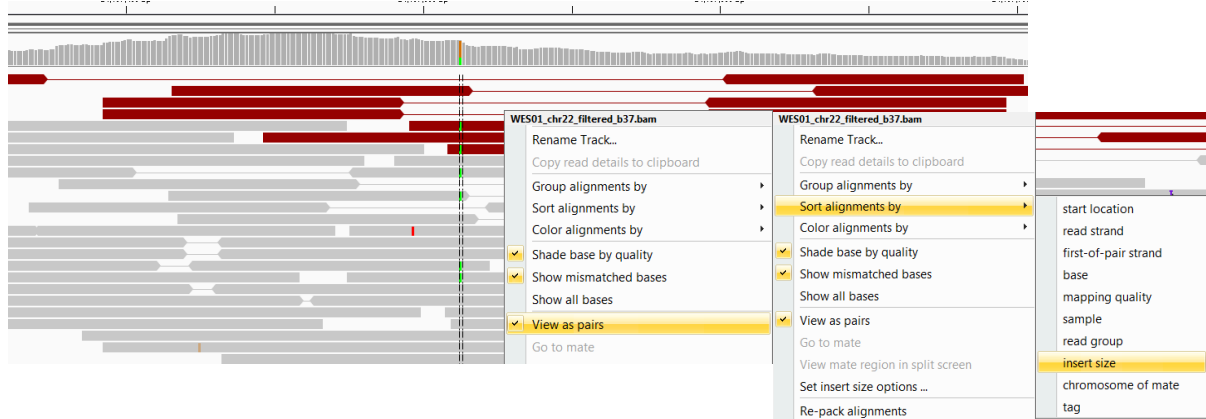
6. Enter the location 'chr22:24,167,513' in the search box to focus on the data for a particular variant. You can now see the coloured bar consists of two colours representing the two alleles and their height corresponds with the allele frequency.

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.