

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 the University of Southampton Galaxy server <http://galaxyngs-hpc.soton.ac.uk> using your iSolutions username and password.

Southampton Logon

Institutional Authentication Gateway





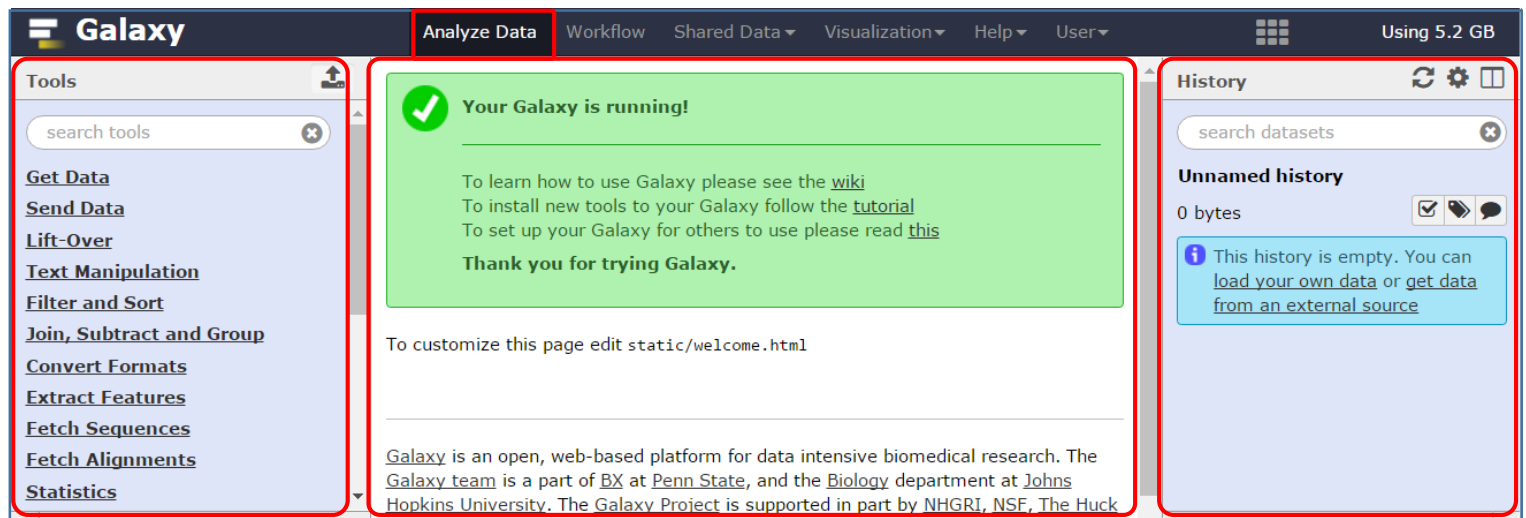
Please enter your University of Southampton user name and password.

Username

Password

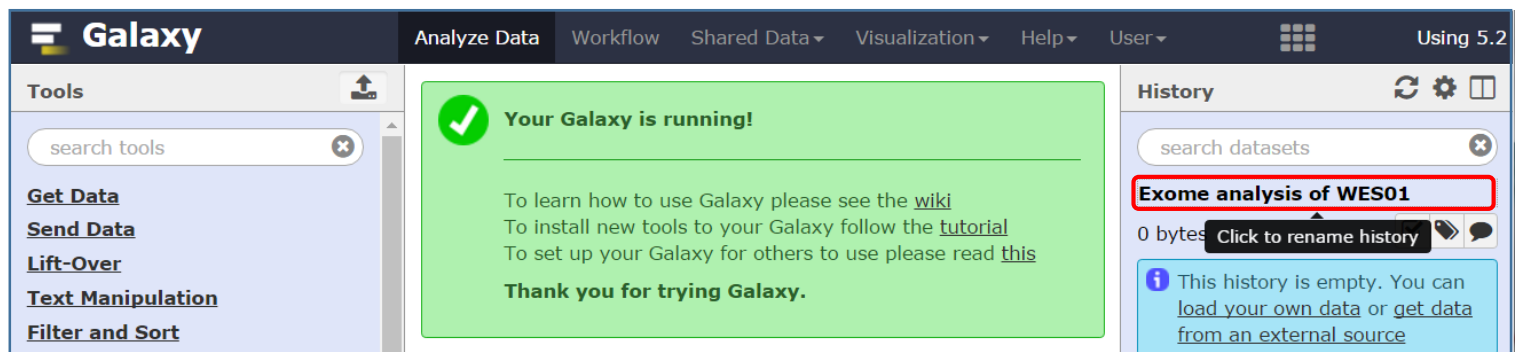
Logon

The Galaxy “Analyse Data” tab consists of three panels. The “Tools” panel on the left hand side gives a list of software for data analysis. The central panel is the working area where you can select tool options, execute jobs and visualise the results. The right hand pane is the “History” which records the tools used and results.



Each Galaxy project should be given its own history with a descriptive name, this will help to organise the results and create workflows.

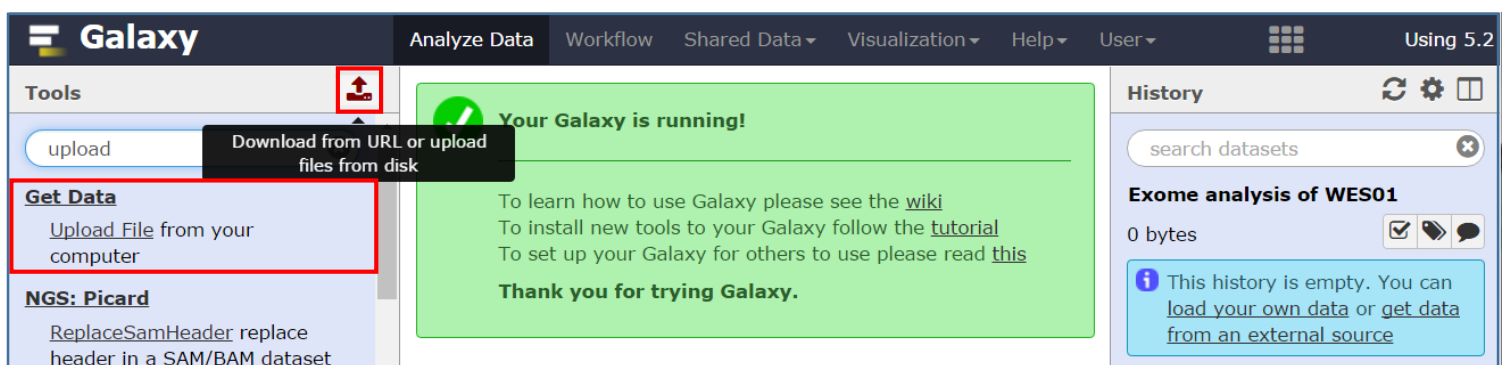
2. Rename the history for this project from “Unnamed history” to “Exome analysis of WES01”. Click on the name, type a new name and press return.



Uploading Files

Upload the raw fastq files from whole exome sequencing to Galaxy using the Upload File tool

1. In the **Tools** pane: Go to **Get Data** > **Upload File** from your computer OR click the upload icon



2. Plug in the USB key and follow steps 1-4 below. Having clicked 'Choose local file' scroll to the bottom and select 'Removable Disk'.

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 5

Download data directly from web or upload files from your disk

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
WES01_chr22_R1.fastq.gz	22 MB	fastqsanger	Human (Homo sap...)		0%
WES01_chr22_R2.fastq.gz	22.3 MB	fastqsanger	Human (Homo sap...)		0%

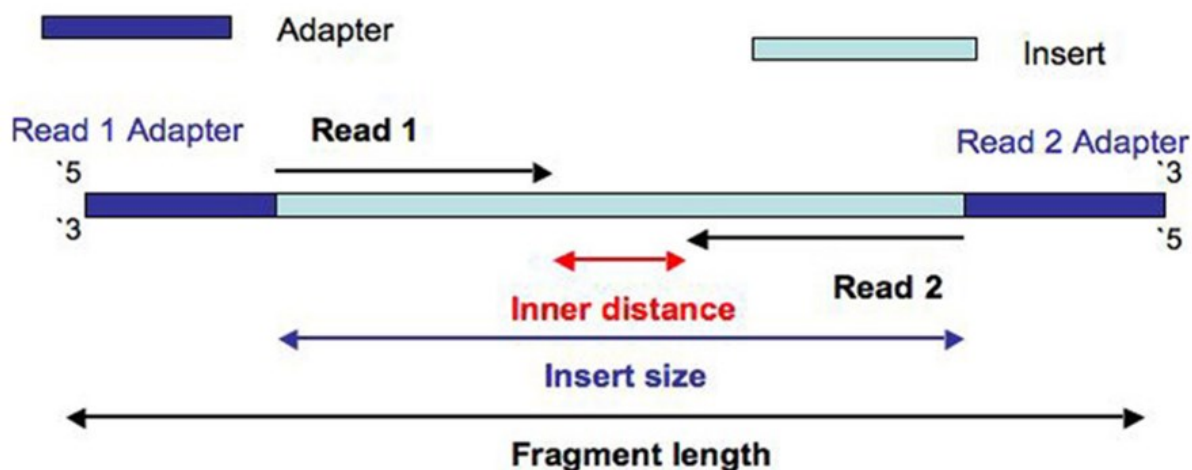
2. Be careful to select fastqsanger and not fastqcsanger
3. Select Human (Homo sapiens) (b37)

Type (set all): fastqsanger Genome (set all): Human (Homo sapiens) (b37)

1 Choose local file 4 Start Pause Reset Close


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**.

The screenshot shows the Galaxy web interface. At the top is a navigation bar with links: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The main content area is divided into three panels. The left panel, titled 'Tools', contains a search bar with 'upload' and a list of tools under 'Get Data' (Upload File from your computer) and 'NGS: Picard' (ReplaceSamHeader, AddCommentsToBam). The center panel displays a green success message: 'Your Galaxy is running!' with instructions on how to use Galaxy and a link to the wiki. The right panel, titled 'History', shows a list of datasets. Two datasets are visible: '2: WES01_chr22_R2.f' and '1: WES01_chr22_R1.f', both with file type 'astq'. Each dataset entry has three icons: an eye (View data), a pencil (Edit attributes), and an 'X' (Delete data). A red box highlights these icons for both datasets. Below the screenshot, three arrows point from the dataset list to a legend: 'View data' points to the eye icon, 'Edit attributes' points to the pencil icon, and 'Delete data' points to the 'X' icon.

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

[illegible]

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

	!"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
Phred score	0.....41
Probability	1.....0.0001
$Phred\ score = -10 \log_{10} P$	

Q1. Use this website (<http://grand-prismatic.blogspot.co.uk/2013/02/fastq-quality-score-convesion-table.html>) and the Sanger sequencing score 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

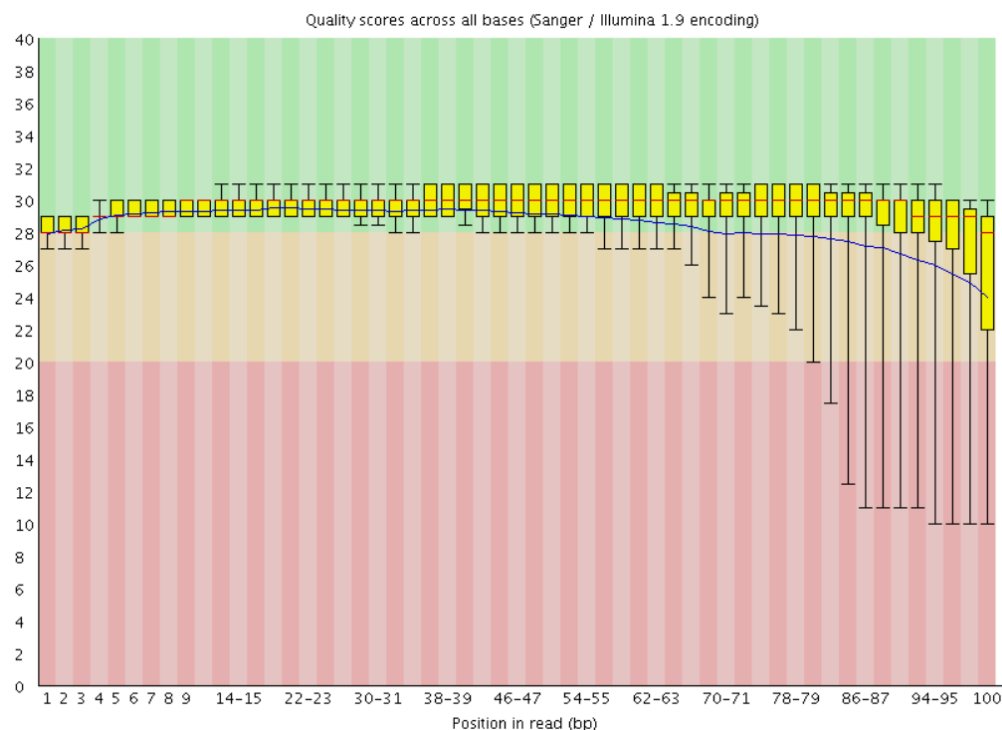
Use the FastQC reports on data 1 and 2 to answer:

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?

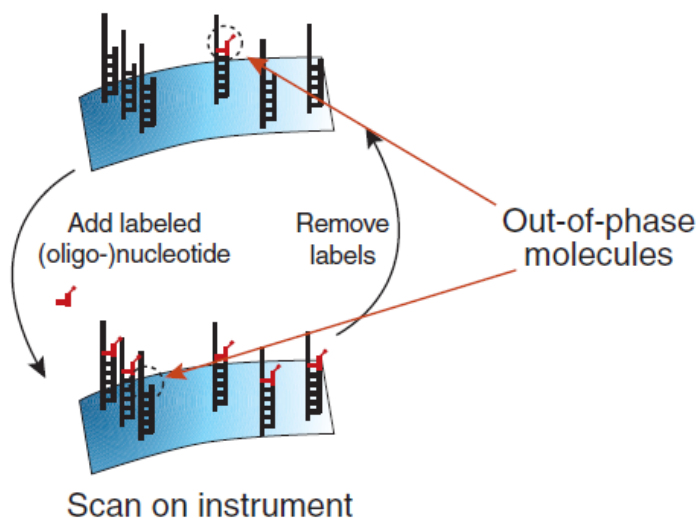
Figure 3. Per base sequence quality for WES01_chr22_R1.fastq



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**

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 5.3 GB

Tools fastq joiner

NGS: QC and manipulation
FASTQ joiner on paired end reads

Workflows
All workflows

FASTQ joiner on paired end reads (Galaxy Tool Version 2.0.0) Versions Options

Left-hand Reads
1: WES01_chr22_R1.fastq

Right-hand Reads
2: WES01_chr22_R2.fastq

FASTQ Header Style
old

Execute

Be careful to select:
WES01_chr22_R1.fastq
WES01_chr22_R2.fastq

History
analysis of WES01
5: FastQC on data 2: W
ebpage
4: FastQC on data 1: R
awData

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 highlighted)

3. Click the link (**8: Filter by quality on data 7**) to expand the job details, then click the 'View details icon' to get more info, finally click the 'stdout' link in the central pane to see the summary results.

The screenshot shows the Galaxy web interface with the job details for '8: Filter by quality on data 7' expanded. The tool details are as follows:

- Tool: Filter by quality**
- Name:** Filter by quality on data 7
- Created:** Tue Jan 24 17:29:35 2017 (UTC)
- Filesize:** 121.3 MB
- Dbkey:** hg_g1k_v37
- Format:** fastqsanger
- Galaxy Tool ID:** toolshed.g2.bx.psu.edu/repos/devteam/fastq_quality_filter/cshl_fastq_quality_filter/1.0.0
- Galaxy Tool Version:** 1.0.0
- Tool Version:**
- Tool Standard stdout:** (link highlighted)
- Tool Standard stderr:**
- Error:**
- Tool Exit Code:** 0
- API ID:** 52d824cf02b980a1
- History ID:** 0fec89153aa8743d
- UUID:** 6aac6c56-f067-40ac-890d-bf4ed1dc7a54

Q5. What number and percentage of reads were discarded because 10% or more bases had Phred quality scores less than 20?

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 highlighted)

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 5.7 GB

Tools

BWA

NGS: SAMtools
MPileup SNP and indel caller

NGS: Mapping
BWA - map short reads (< 100 bp) against reference genome
BWA-MEM - map medium and long reads (> 100 bp) against reference genome

NGS: Picard
FilterSamReads include or exclude aligned and unaligned reads and read lists

Workflows
All workflows

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

genome (Galaxy Tool Version 0.1)

Load reference genome from
Local cache

Using reference genome
Human (Homo sapiens) (b37)
Select genome from the list

Single or Paired-end reads
Paired
Select between paired and single end data

Select first set of reads
9: FASTQ splitter on data 8
Specify dataset with forward reads

Select second set of reads
10: FASTQ splitter on data 8
Specify dataset with reverse reads

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.

Be careful to select:
9: FASTQ splitter on data 8
10: FASTQ splitter on data 8

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

Execute

History
search datasets

Exome analysis of WES01
14 shown
581.3 MB

14: FastQC on data 10: RawData
13: FastQC on data 10: Webpage
12: FastQC on data 9: RawData
7: FASTQ joiner on data 2 and data 1
6: FastQC on data 2:

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:

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 provides simple stats on BAM files

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

NGS: SAMtools
Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region

NGS: Picard
EstimateLibraryComplexity assess sequence library complexity from read sequences
FilterSamReads include or exclude aligned and unaligned reads and read lists

Workflows
All workflows

Skip alignments with any of these flag bits set
Select/Unselect all
☐ 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
 (-F)

Select the output format
BAM (-b)
Execute

Exome analysis of WES01
16 shown
620.2 MB

13: FastQC on data
10: Webpage
12: FastQC on data
9: RawData

10: FASTQ splitter on data 8
9: FASTQ splitter on data 8

2. 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 Help User Using 5.8 GB

Tools
flagstat

NGS: SAMtools
flagstat provides simple stats on BAM files

flagstat provides simple stats on BAM files (Galaxy Tool Version 1.0.0)
Options

BAM File to Convert
17: Filter SAM or BAM, output SAM or BAM on data 15: bam
Execute

History
search datasets
Exome analysis of WES01
17 shown
657.5 MB
18: flagstat on data 17
17: Filter SAM or BAM, output SAM or BAM on data 15: bam

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

Galaxy
Analyze Data Workflow Shared Data Visualization Help User Using 5.8 GB

Tools
flagstat

NGS: SAMtools
flagstat provides simple stats on BAM files

Workflows
All workflows

1 job has been successfully added to the queue - resulting in the following datasets:
18: flagstat on data 17
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
Exome 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
View data

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.

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:

- ### Table 1. SAM file format

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.

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)

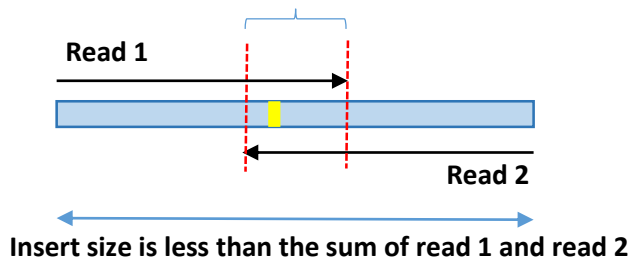
1	2	3	4
MT	16569	0	0
1	249250621	26	0
2	243199373	31	0
3	198022430	10	0
4	191154276	21	0
5	180915260	18	0
6	171115067	7	0
7	159138663	18	0
8	146364022	32	0
9	141213431	9	0
10	135534747	40	0

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.

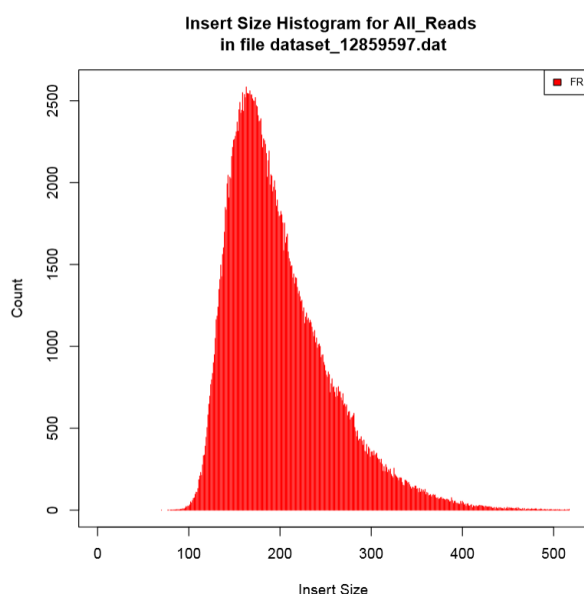
1. In **Tool Pane**: Go to **NGS: Picard** > CollectInsertSizeMetrics

Select the filtered bam file, Human hg19 reference genome, don't change the other settings and click execute.

This tool produces 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' pane lists 'NGS: Picard' and 'AddOrReplaceReadGroups add or replaces read group information'. The main tool pane shows the configuration for 'AddOrReplaceReadGroups' (Tool Version 1.126.0). 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 'tumor-a'. The 'Select validation stringency' is set to 'Lenient'. The 'Execute' button is highlighted. The right sidebar shows a history of datasets, including 'Exome analysis of WES01' and '22: CollectInsertSizeMetrics on data 17'.

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 'Upload File' interface. The 'Download data directly from web or upload files from your disk' section shows a table with one file: '22_agilent50_targets_hg19.bed' (92.9 KB, bed type, unspecified genome). The 'Type (set all):' dropdown is set to 'bed' and the 'Genome (set all):' dropdown is set to 'Human (Homo sapiens) (b37)'. The 'Start' button is highlighted.

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 left sidebar shows the 'Tools' section with 'NGS: GATK Tools' and 'Depth of Coverage on BAM files' highlighted. The main configuration area is titled 'Depth of Coverage on BAM files (Galaxy Tool Version 0.0.2)'. It includes sections for 'Choose the source for the reference list', 'BAM file', 'Using reference genome', 'Output format', 'Basic or Advanced GATK options', 'Operate on Genomic intervals', and 'Basic or Advanced Analysis options'. Several elements are highlighted with red boxes: 'NGS: GATK Tools', 'Depth of Coverage on BAM files', '23: AddOrReplaceReadGroups on data 17: BAM with replaced/...', 'Human (Homo sapiens) (b37)', 'table' for output format, 'Advanced' for GATK options, '24: 22_agilent50_targets_hg19.bed' for genomic intervals, and the 'Execute' button. The right panel shows a history of previous runs, including 'Exome analysis of WES01' and various intermediate steps like '24: 22_agilent50_targets_hg19.bed', '23: AddOrReplaceReadGroups on data 17: BAM with replaced/modified readgroups', '22: CollectInsertSizeMetrics on data 17', '21: CollectInsertSizeMetrics on data 17', '20: IdxStats on data 17', '19: BAM-to-SAM on data 17: converted', '18: flagstat on data 17', '19: BAM-to-SAM on data 17: converted SAM', and '18: flagstat on data 17'.

Q11. Use the output summary sample result to determine the mean coverage and percentage of target bases covered by 15 or more reads according to the depth of coverage output.

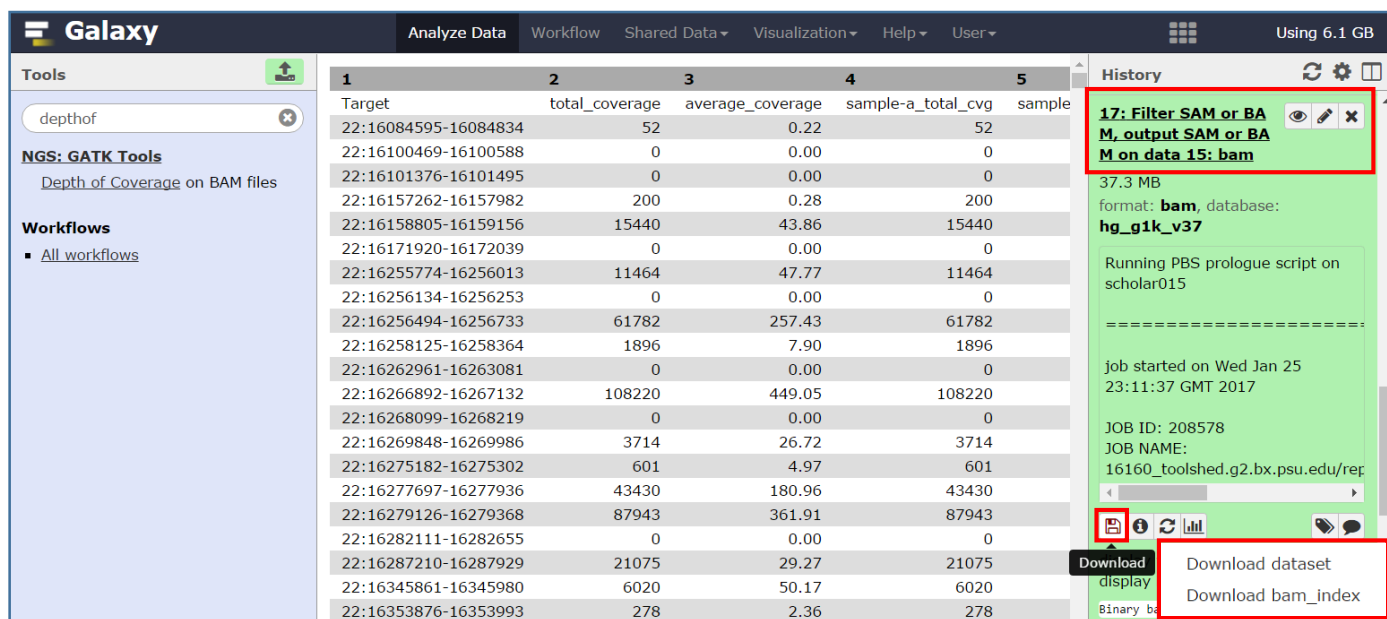
Depth of coverage produces several other output files which include statistics broken down by intervals.

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:

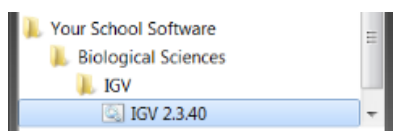
1. Download the dataset (BAM file) and the bam_index file from step no. 17



The screenshot shows the Galaxy web interface. On the left, there's a 'Tools' panel with 'NGS: GATK Tools' and 'Workflows'. The main panel displays a table with 5 columns: Target, total_coverage, average_coverage, sample-a_total_cvg, and sample. The table contains genomic data for various targets. On the right, the 'History' panel shows a job titled '17: Filter SAM or BAM, output SAM or BAM on data 15: bam'. Below the job title, it shows the format as 'bam', database as 'hg_g1k_v37', and the job started on Wed Jan 25 23:11:37 GMT 2017. At the bottom of the history panel, there are buttons for 'Download dataset' and 'Download bam_index'.

1	2	3	4	5
Target	total_coverage	average_coverage	sample-a_total_cvg	sample
22:16084595-16084834	52	0.22	52	
22:16100469-16100588	0	0.00	0	
22:16101376-16101495	0	0.00	0	
22:16157262-16157982	200	0.28	200	
22:16158805-16159156	15440	43.86	15440	
22:16171920-16172039	0	0.00	0	
22:16255774-16256013	11464	47.77	11464	
22:16256134-16256253	0	0.00	0	
22:16256494-16256733	61782	257.43	61782	
22:16258125-16258364	1896	7.90	1896	
22:16262961-16263081	0	0.00	0	
22:16266892-16267132	108220	449.05	108220	
22:16268099-16268219	0	0.00	0	
22:16269848-16269986	3714	26.72	3714	
22:16275182-16275302	601	4.97	601	
22:16277697-16277936	43430	180.96	43430	
22:16279126-16279368	87943	361.91	87943	
22:16282111-16282655	0	0.00	0	
22:16287210-16287929	21075	29.27	21075	
22:16345861-16345980	6020	50.17	6020	
22:16353876-16353993	278	2.36	278	

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

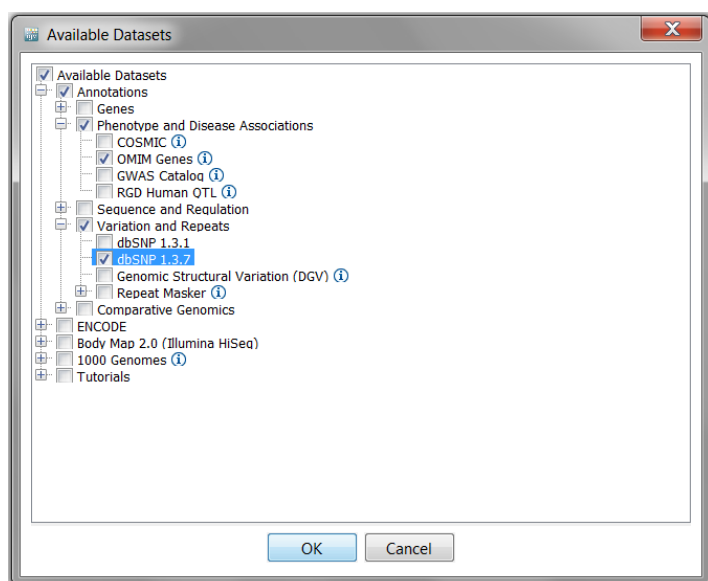
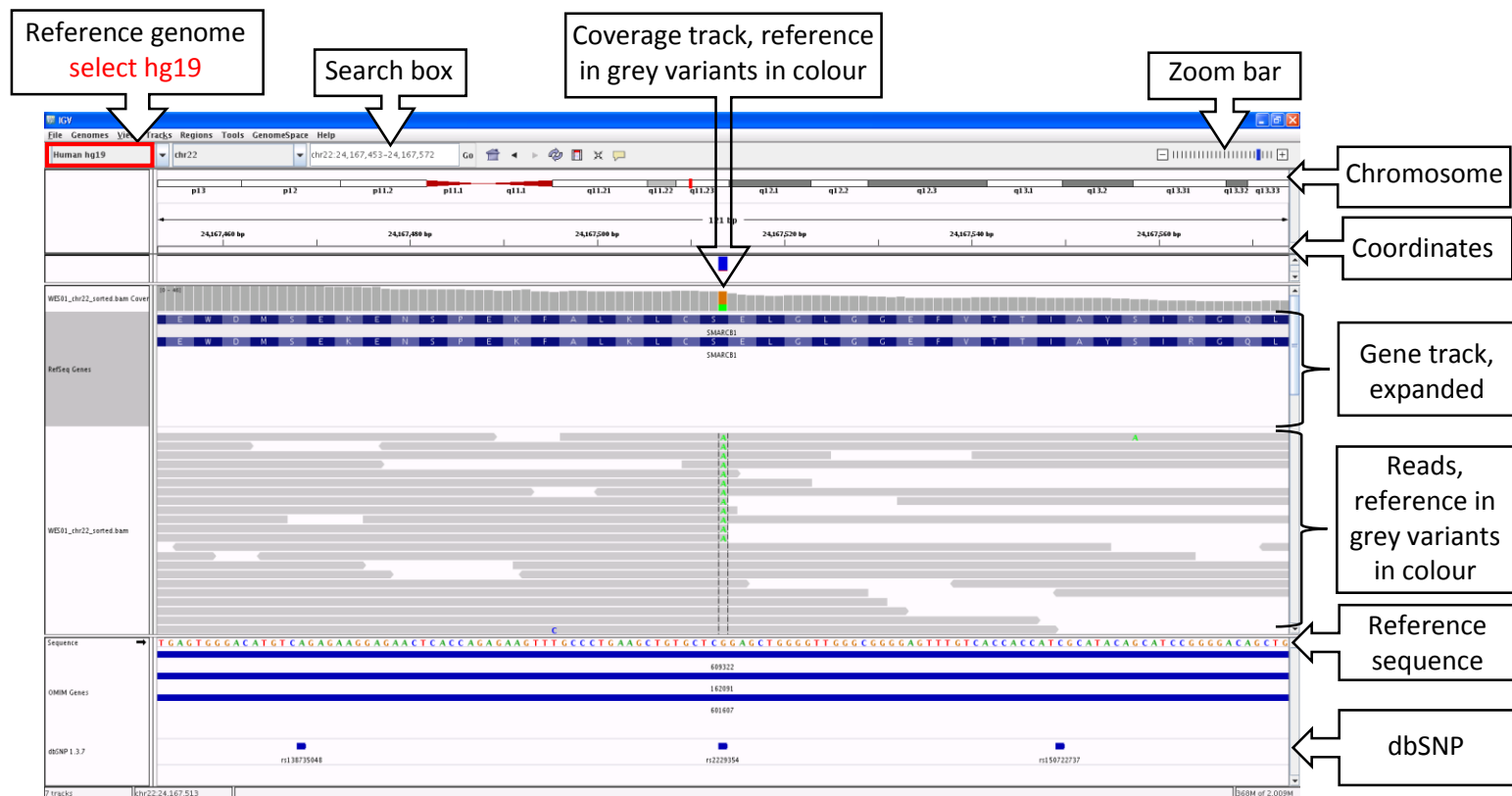


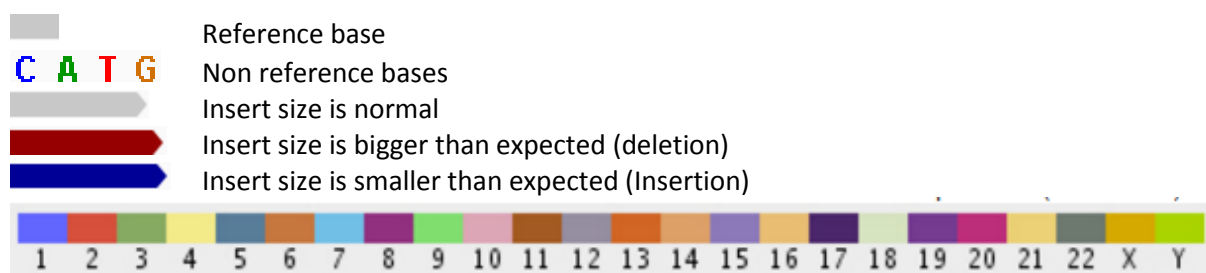
Figure 6. The Integrative Genome Viewer (IGV)

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.

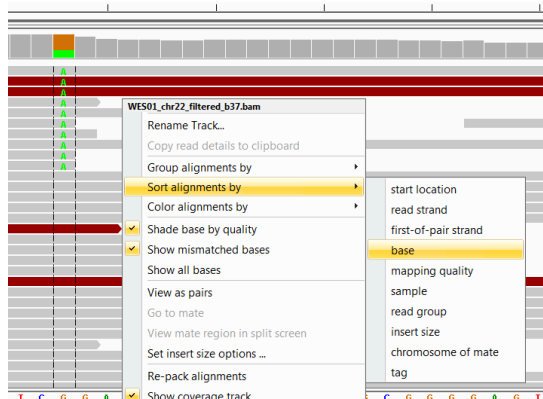
Look here http://www.broadinstitute.org/software/igv/interpreting_insert_size for more details on the colour codes and settings.

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.

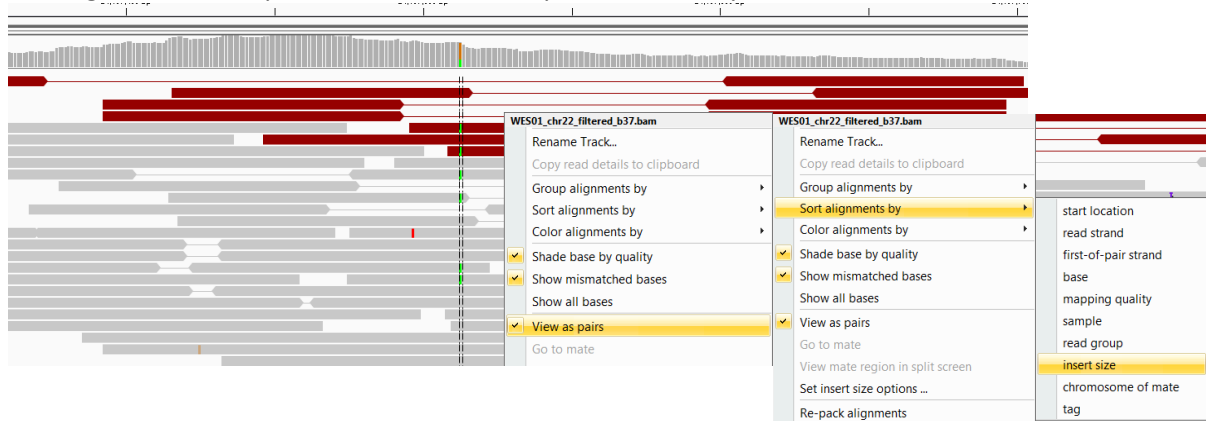
6. Enter the location 'chr22:24,167,513' in the search box to focus on the data for a particular variant. You can now see the coloured bar consists of two colours representing the two alleles and their height corresponds with the allele frequency.

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.