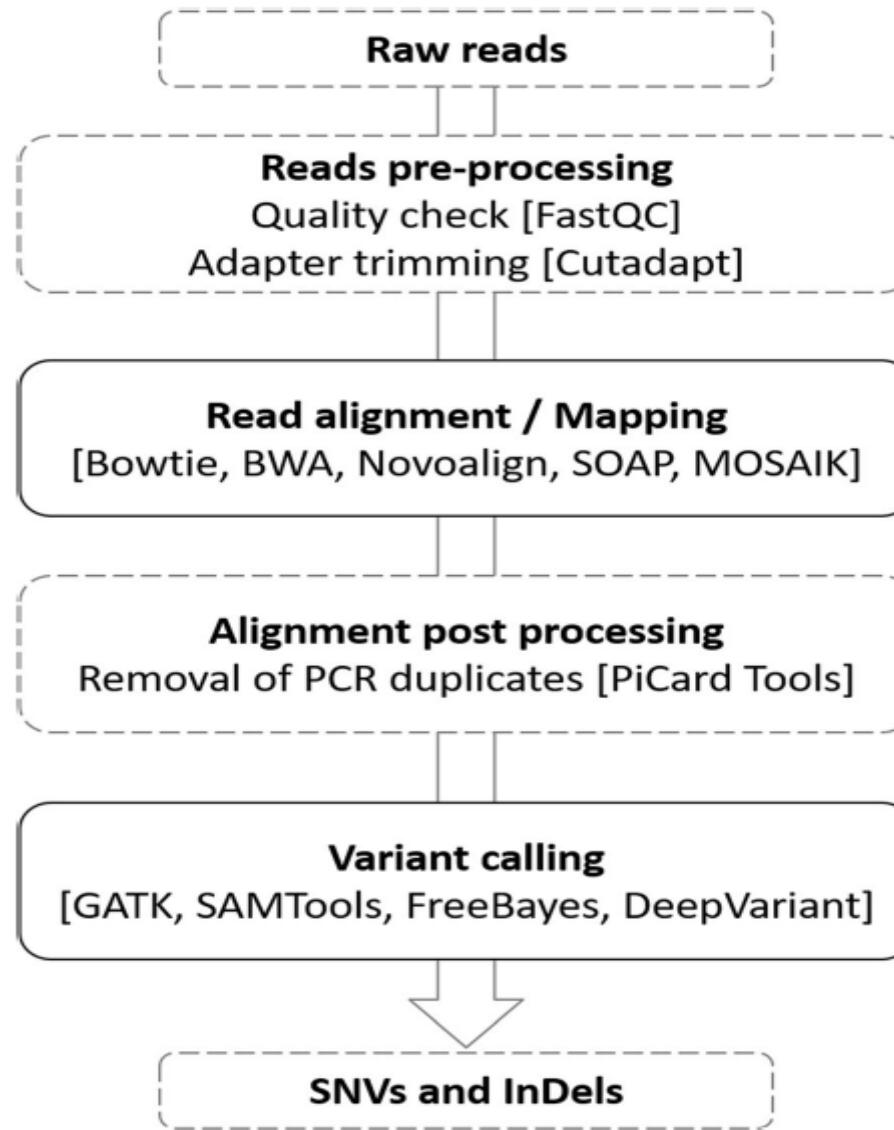


# HANDS ON: INTRODUCTION TO VARIANT ANALYSIS

Bioinformàtica per a la Recerca Biomèdica  
**Ricardo Gonzalo Sanz**  
[ricardo.gonzalo@vhir.org](mailto:ricardo.gonzalo@vhir.org)



This presentation is mainly based in a tutorial published by:



**Melbourne Bioinformatics**  
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

## Tutorials and protocols

These tutorials have been developed by bioinformaticians at Melbourne Bioinformatics (formerly VLSCI). They are regularly delivered on-site or may be run in-house for your group. training materials were developed for use on the Australian-made Genomics Virtual Lab and these are used in our formal workshops and are also available for use to deliver workshops or for self-directed learning.

<https://www.melbournebioinformatics.org.au/tutorials/>



## Introduction to Variant Calling using Galaxy

[https://www.melbournebioinformatics.org.au/tutorials/tutorials/variant\\_calling\\_galaxy\\_1/variant\\_calling\\_galaxy\\_1/](https://www.melbournebioinformatics.org.au/tutorials/tutorials/variant_calling_galaxy_1/variant_calling_galaxy_1/)



## Outline:

1. Load the data
2. Aligns the reads to the genome. QC of the data
3. Look for differences between reads and reference genome sequence
4. Visualise BAM files using Genomics Viewer
5. Detect small variants (SNVs and indels)
6. Filter the detected genomic variation
7. Annotate the detected genomic variation

## Objective:

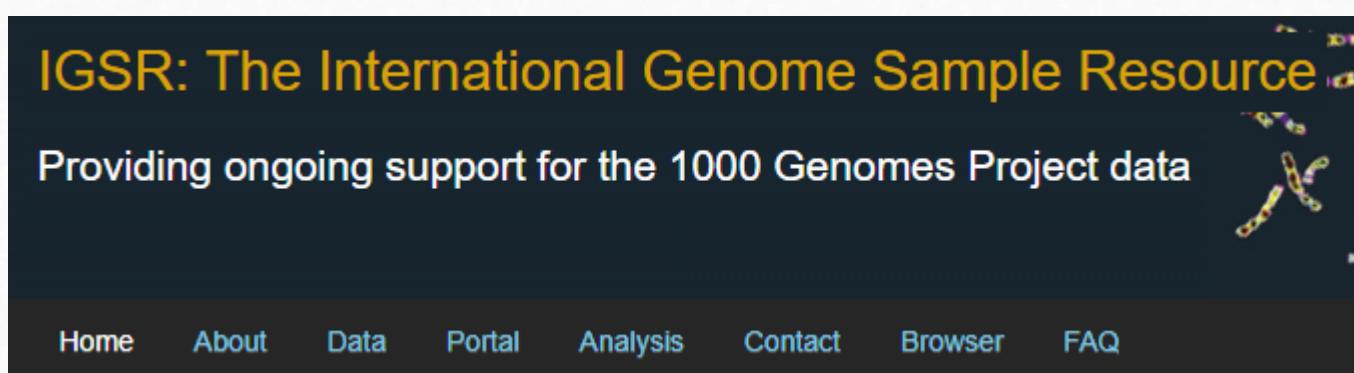
Detecting small variants in human genomic DNA using a small set of reads from chromosome 22

## Data

Analysis of short read data from the exome of Chr 22 of a single human individual.

There are **one million of 76bp reads** in the dataset produced on an Illumina GAIIx.

Data generated as part of the 1000 genomes project



The image shows a screenshot of the International Genome Sample Resource (IGSR) website. The header features the text "IGSR: The International Genome Sample Resource" in yellow, followed by "Providing ongoing support for the 1000 Genomes Project data" in white. A decorative graphic of chromosomes is visible on the right side. At the bottom, a dark navigation bar contains links for "Home", "About", "Data", "Portal", "Analysis", "Contact", "Browser", and "FAQ".

1. Open Galaxy and create a new history and name it ("Exon Variant Analysis")

<https://usegalaxy.eu/>

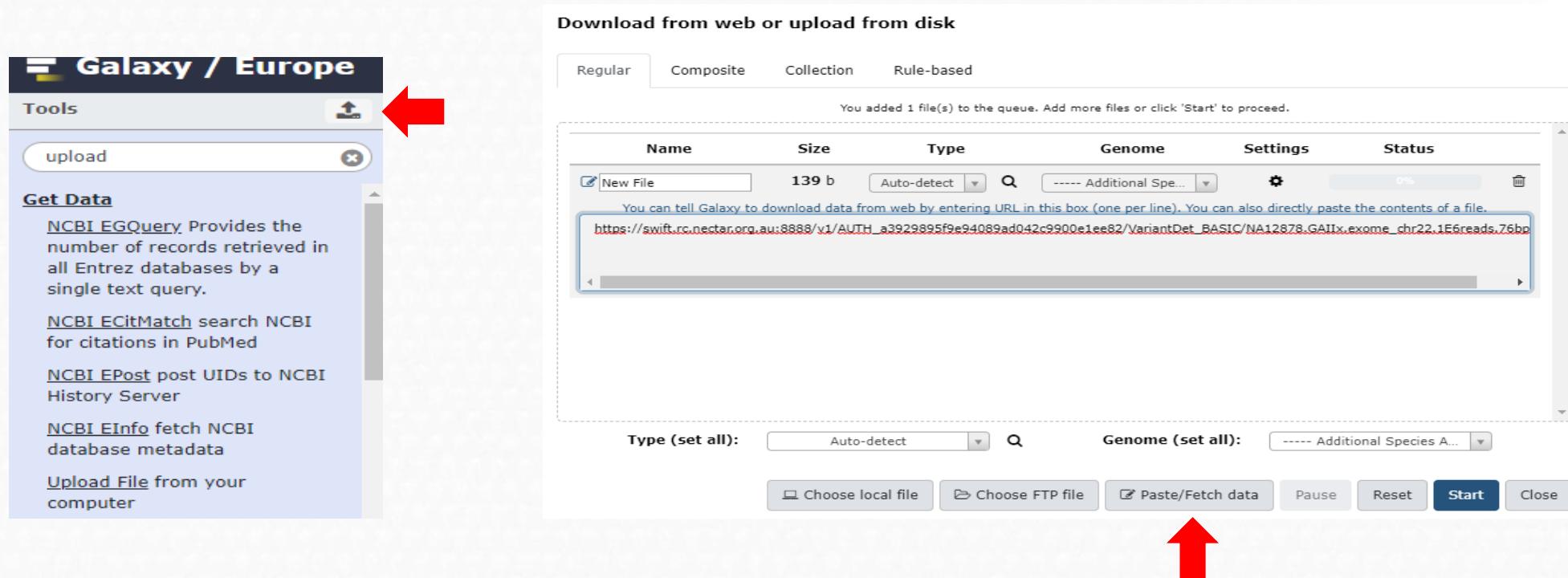
[https://usegalaxy.eu/join-training/ueb\\_bi2020](https://usegalaxy.eu/join-training/ueb_bi2020)



## Import data for the tutorial

1. Paste the following link to download the data

[https://swift.rc.nectar.org.au:8888/v1/AUTH\\_a3929895f9e94089ad042c9900e1ee82/VariantDet\\_BASIC/NA12878.GAIix.exome\\_chr22.1E6reads.76bp.fastq](https://swift.rc.nectar.org.au:8888/v1/AUTH_a3929895f9e94089ad042c9900e1ee82/VariantDet_BASIC/NA12878.GAIix.exome_chr22.1E6reads.76bp.fastq)



The screenshot shows the Galaxy / Europe interface. On the left, there's a sidebar with 'Tools' and several options: 'upload' (highlighted with a red arrow), 'Get Data' (with sub-options like NCBI EGQuery, ECitMatch, EPost, EInfo, and Upload File from your computer), and 'Upload File from your computer'.

The main area is titled 'Download from web or upload from disk'. It shows a table with one item:

Name	Size	Type	Genome	Settings	Status
New File	139 b	Auto-detect	----- Additional Spec...		0%

A text input field contains the URL: [https://swift.rc.nectar.org.au:8888/v1/AUTH\\_a3929895f9e94089ad042c9900e1ee82/VariantDet\\_BASIC/NA12878.GAIix.exome\\_chr22.1E6reads.76bp.fastq](https://swift.rc.nectar.org.au:8888/v1/AUTH_a3929895f9e94089ad042c9900e1ee82/VariantDet_BASIC/NA12878.GAIix.exome_chr22.1E6reads.76bp.fastq). A tooltip above the input field explains: 'You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.'

At the bottom, there are buttons for 'Type (set all)', 'Genome (set all)', and file selection ('Choose local file', 'Choose FTP file', 'Paste/Fetch data'). The 'Paste/Fetch data' button is highlighted with a red arrow. Other buttons include 'Pause', 'Reset', 'Start', and 'Close'.

Change the name of file:

### Exon Variant Analysis

1 shown

209.99 MB



1: [https://swift.rc.nectar.org.au:8888/v1/AUTH\\_a3929895f9e94089ad042c9900e1ee8/VariantDet\\_BASIC/NA12878.GAIix.exome\\_chr22.1E6reads.76bp.fastq](https://swift.rc.nectar.org.au:8888/v1/AUTH_a3929895f9e94089ad042c9900e1ee8/VariantDet_BASIC/NA12878.GAIix.exome_chr22.1E6reads.76bp.fastq)



### Edit dataset attributes

Attributes  Convert  Datatypes  Permissions

Edit attributes

Auto-detect  Save

#### Name

NA12878.fastq

#### Info

uploaded fastqsanger file

### Exon Variant Analysis

1 shown

209.99 MB



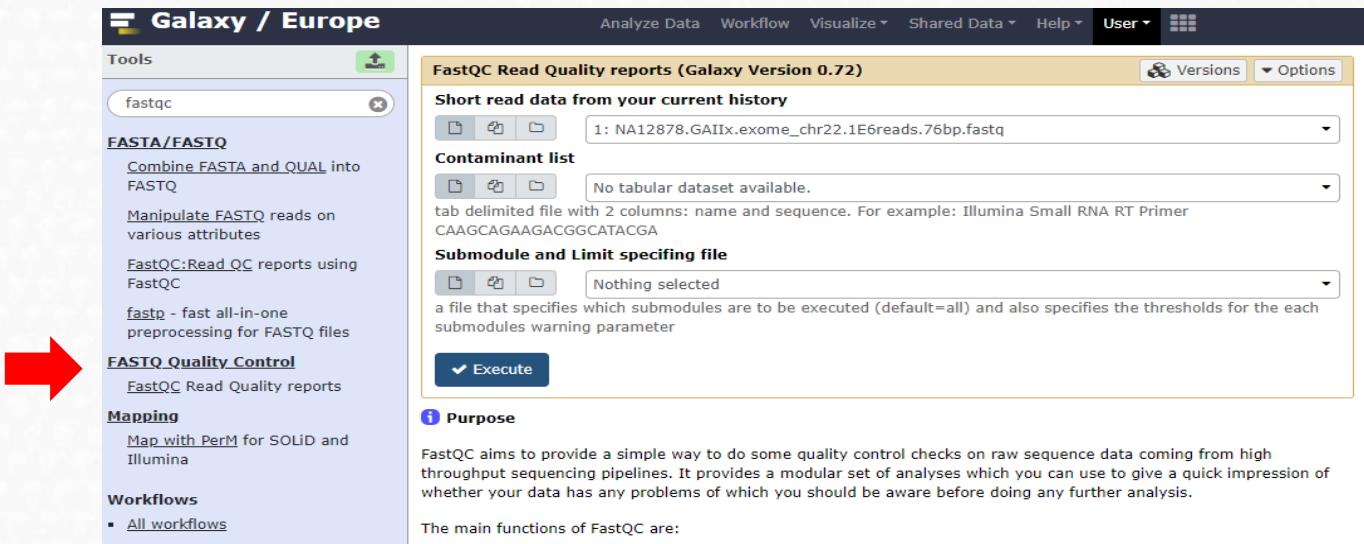
1: NA12878.fastq



- Take a look at the FASTQ file

```
@61CC3AAXX100125:7:118:2538:5577/1
GACACCTTTAATGTCTGAAAAAGAGACATTACCATCTATTCTCTTGGAGGGCTACCACCTAAGAGCCTTCATCCCC
+
?>CADFEEEEDIEHHIDGGGESEEHFFGIGIIFFIIEFHIIIIHIIFFIIIDEIIIGIIIIEHFFFIIIEHIFAO?==
```

### 3. Assessing read quality from the FASTQ files



The screenshot shows the Galaxy Europe interface. On the left, a sidebar lists various tools and workflows. A red arrow points to the "fastqc" entry under the "Tools" section. The main workspace is titled "FastQC Read Quality reports (Galaxy Version 0.72)". It contains several configuration sections: "Short read data from your current history" (with a dropdown menu showing "1: NA12878.GAIIX.exome\_chr22.1E6reads.76bp.fastq"), "Contaminant list" (with a dropdown menu showing "No tabular dataset available."), "Submodule and Limit specifying file" (with a dropdown menu showing "Nothing selected"), and a "Purpose" section with a detailed description of FastQC's function. At the bottom of the workspace, there is a note about the main functions of FastQC.

**Galaxy / Europe**

Analyze Data Workflow Visualize Shared Data Help User

**Tools**

fastqc

**FASTA/FASTQ**

Combine FASTA and QUAL into FASTQ

Manipulate FASTQ reads on various attributes

FastQC:Read\_QC reports using FastQC

fastp - fast all-in-one preprocessing for FASTQ files

**FASTQ Quality Control**

FastQC Read Quality reports

**Mapping**

Map with PerM for SOLiD and Illumina

**Workflows**

All workflows

**FastQC Read Quality reports (Galaxy Version 0.72)**

Versions Options

Short read data from your current history

1: NA12878.GAIIX.exome\_chr22.1E6reads.76bp.fastq

Contaminant list

No tabular dataset available.

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAAGACGGCATACGA

Submodule and Limit specifying file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for each submodules warning parameter

Execute

**Purpose**

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Results of the Quality Control



OR



## 4. Align the reads with BWA

Reference Genome: *Human genome 19 (hg19)*

**Galaxy / Australia**

Analyze Data Workflow Visualize Shared Data Help User

**Tools**

- bwa

**FASTQ Quality Control**

- Trim Galore! Quality and adapter trimmer of reads

**SAM/BAM**

- Filter SAM on bitwise flag values

**Annotation**

- SnpEff build: database from Genbank or GFF record Updated

**Mapping**

- Map with minimap2 A fast pairwise aligner for genomic and spliced nucleotide sequences
- Map with BWA-MEM** - map medium and long reads (> 100 bp) against reference genome
- Map with BWA - map short reads (< 100 bp) against reference genome

**Variant Calling**

- snippy Snippy finds SNPs between a haploid reference genome and your NGS sequence reads.

**Picard**

- FilterSamReads include or exclude aligned and unaligned reads and read lists

**Workflows**

**Map with BWA-MEM** - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.1)

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See "Indexes" section of help below

Using reference genome

Human (Homo sapiens) (37): hg19

Select genome from the list

Single or Paired-end reads

Single

Select between paired and single end data

Select fastq dataset

1: fastq

Specify dataset with single reads

Set read groups information?

Do not set

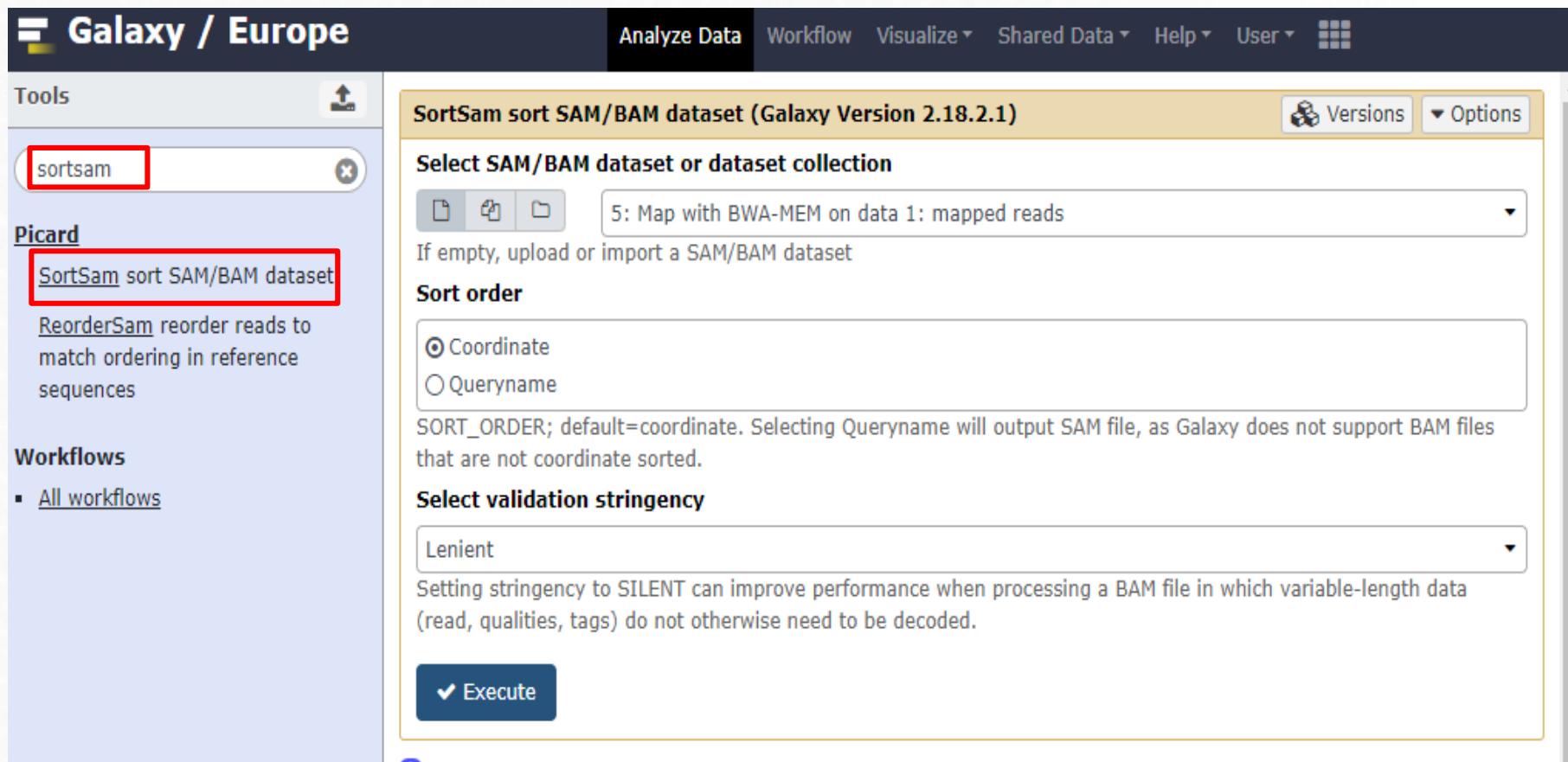
Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

1.Simple Illumina mode

Execute

- Sort the BAM file

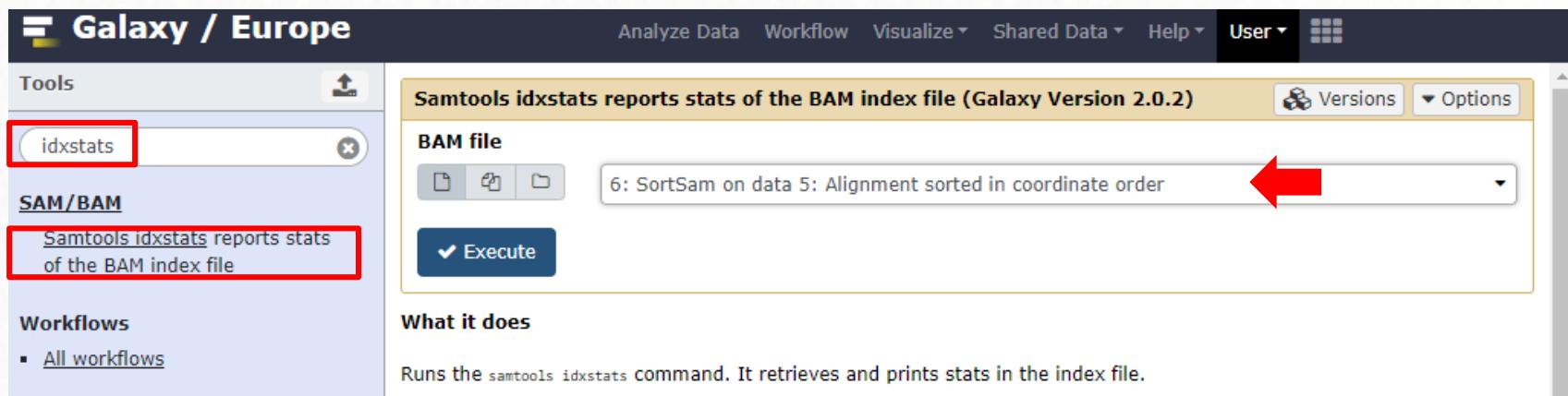


The screenshot shows the Galaxy Europe web interface. The top navigation bar includes "Analyze Data", "Workflow", "Visualize", "Shared Data", "Help", "User", and a grid icon. On the left, a sidebar titled "Tools" lists "sortsam" under "Picard". Below it, "Workflows" is expanded, showing "All workflows". The main content area is titled "SortSam sort SAM/BAM dataset (Galaxy Version 2.18.2.1)". It features a search bar with "sortsam" and a dropdown menu showing "5: Map with BWA-MEM on data 1: mapped reads". Under "Sort order", "Coordinate" is selected. A note states: "SORT\_ORDER; default=coordinate. Selecting Queryname will output SAM file, as Galaxy does not support BAM files that are not coordinate sorted." The "Select validation stringency" section has "Lenient" selected. A note says: "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." At the bottom is a blue "Execute" button.

61CC3AAXX100125://118:2538:55//	16	chr22	16050954
61CC3AAXX100125:7:1:17320:13701	0	chr22	16052274
61CC3AAXX100125:7:93:5100:14497	0	chr14	19790076
61CC3AAXX100125:6:92:7549:15004	16	chr22	16052936
61CC3AAXX100125:5:7:1488:7780	16	chr22	16053177
61CC3AAXX100125:7:72:14903:20386	16	chr22	16053702
61CC3AAXX100125:7:88:9942:19183	0	chr14	19788896
61CC3AAXX100125:7:76:1585:2024	0	chr22	16054020
61CC3AAXX100125:6:26:17654:5573	0	chr22	16053945
61CC3AAXX100125:7:117:7805:10957	0	chr14	19788482
61CC3AAXX100125:7:36:11248:16392	0	chr22	16054533
61CC3AAXX100125:6:80:10088:8830	16	chr22	16054924
61CC3AAXX100125:6:115:5701:20053	0	chr22	16055354
61CC3AAXX100125:5:20:10205:7274	0	chr14	19787528
61CC3AAXX100125:6:22:16350:6073	16	chr14	19787506
61CC3AAXX100125:7:120:16647:15768	16	chr14	19787513
61CC3AAXX100125:7:107:14497:1691	16	chr22	16055582
61CC3AAXX100125:5:71:19423:10946	16	chr22	16055583
61CC3AAXX100125:5:103:9987:17912	16	chr22	16055617
61CC3AAXX100125:6:33:7020:21084	0	chr14	19787108
61CC3AAXX100125:7:71:19300:18871	0	chr22	16055656
61CC3AAXX100125:6:37:4641:21236	0	chr22	16056042
61CC3AAXX100125:7:1:15981:6383	16	chr22	16056227
61CC3AAXX100125:6:74:11878:18737	0	chr22	16056323
61CC3AAXX100125:7:50:6601:7254	0	chr22	16056435
61CC3AAXX100125:7:38:15573:6120	16	chr14	19786389
61CC3AAXX100125:6:95:9677:19470	0	chr22	16057096

the aligner does the best it can, but because of compromises in accuracy vs performance and repetitive sequences in the genome, not all the reads will necessarily align to the 'correct' sequence

## 5. Assess the alignment data



The screenshot shows the Galaxy Europe interface. On the left, the 'Tools' panel is open, displaying the 'idxstats' tool under the 'SAM/BAM' category. A red box highlights the 'idxstats' entry. Below it, a red box highlights the descriptive text: 'Samtools idxstats reports stats of the BAM index file'. The main workspace shows a workflow step titled 'Samtools idxstats reports stats of the BAM index file'. This step is part of a larger workflow: '6: SortSam on data 5: Alignment sorted in coordinate order'. A red arrow points to this step. The 'Execute' button is visible below the step title. To the right, the 'What it does' section describes the tool: 'Runs the `samtools idxstats` command. It retrieves and prints stats in the index file.'

generate some mapping statistics from the BAM file

1	2	3	4
chr10	135534747	786	0
chr11	135006516	5541	0
chr11_gl000202_random	40103	0	0
chr12	133851895	534	0
chr13	115169878	199	0
chr14	107349540	12913	0
chr15	102531392	621	0
chr16	90354753	724	0
chr17_ctg5_hap1	1680828	2	0
chr17	81195210	275	0
chr17_gl000203_random	37498	0	0
chr17_gl000204_random	81310	0	0
chr17_gl000205_random	174588	0	0
chr17_gl000206_random	41001	0	0
chr18	78077248	474	0
chr18_gl000207_random	4262	0	0
chr19	59128983	434	0
chr19_gl000208_random	92689	0	0
chr19_gl000209_random	159169	0	0
chr1	249250621	1106	0
chr1_gl000191_random	106433	0	0
chr1_gl000192_random	547496	3	0
chr20	63025520	281	0
chr21	48129895	573	0
chr21_gl000210_random	27682	0	0
chr22	51304566	1101584	0
chr2	243199373	4894	0
chr3	198022430	1099	0
chr4_ctg9_hap1	590426	0	0
chr4	191154276	663	0
chr4_gl000193_random	189789	0	0
chr4_gl000194_random	191469	20	0
chr5	180915260	234	0
chr6_apd_hap1	4622290	0	0
chr6_cox_hap2	4795371	13	0
chr7	161111111	15	0

## Column Description

-----

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)

## 6. Visualise the BAM file with IGV

5: SortSam on data 4: Alignme  
nt sorted in coordinate order

64.3 MB

format: **bam**, database: **hg19**

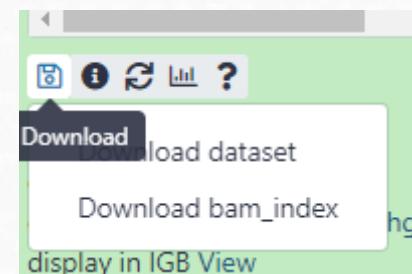
```
Picked up _JAVA_OPTIONS: -Xmx4G -Xms1G
-Djava.io.tmpdir=/data/2/galaxy_db/tmp
17:55:47.372 INFO NativeLibraryLoader -
Loading libgkl_compression.so from
jar:file:/usr/local/tools/_conda/envs/_picard@2.18.2-0/picard.jar!/com/intel/gk
```

◀ ▶

?

play at UCSC main  
 display at Ensembl Current  
 display with IGV local Human hg19  
 display in IGB View  
 display at bam.iobio.bam.iobio.io

Binary bam alignments file



<https://software.broadinstitute.org/software/igv/download>

Home > Downloads

## Downloads

Did you know that there is also an IGV web application that runs only in a web browser, does not use Java, and requires no downloads? See <https://igv.org/app>. Click on the [Help](#) link in the app for more information about using IGV-Web.

### Install IGV 2.8.13

See the [Release Notes](#) for what's new in each IGV release.



IGV MacOS App  
Java included



IGV MacOS App  
Separate Java 11 required



IGV for Windows  
Java included



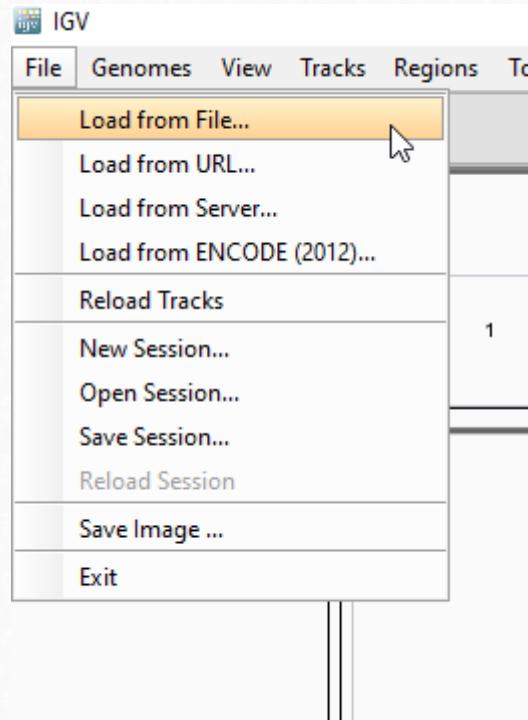
IGV for Windows  
Separate Java 11 required



IGV for Linux  
Java included

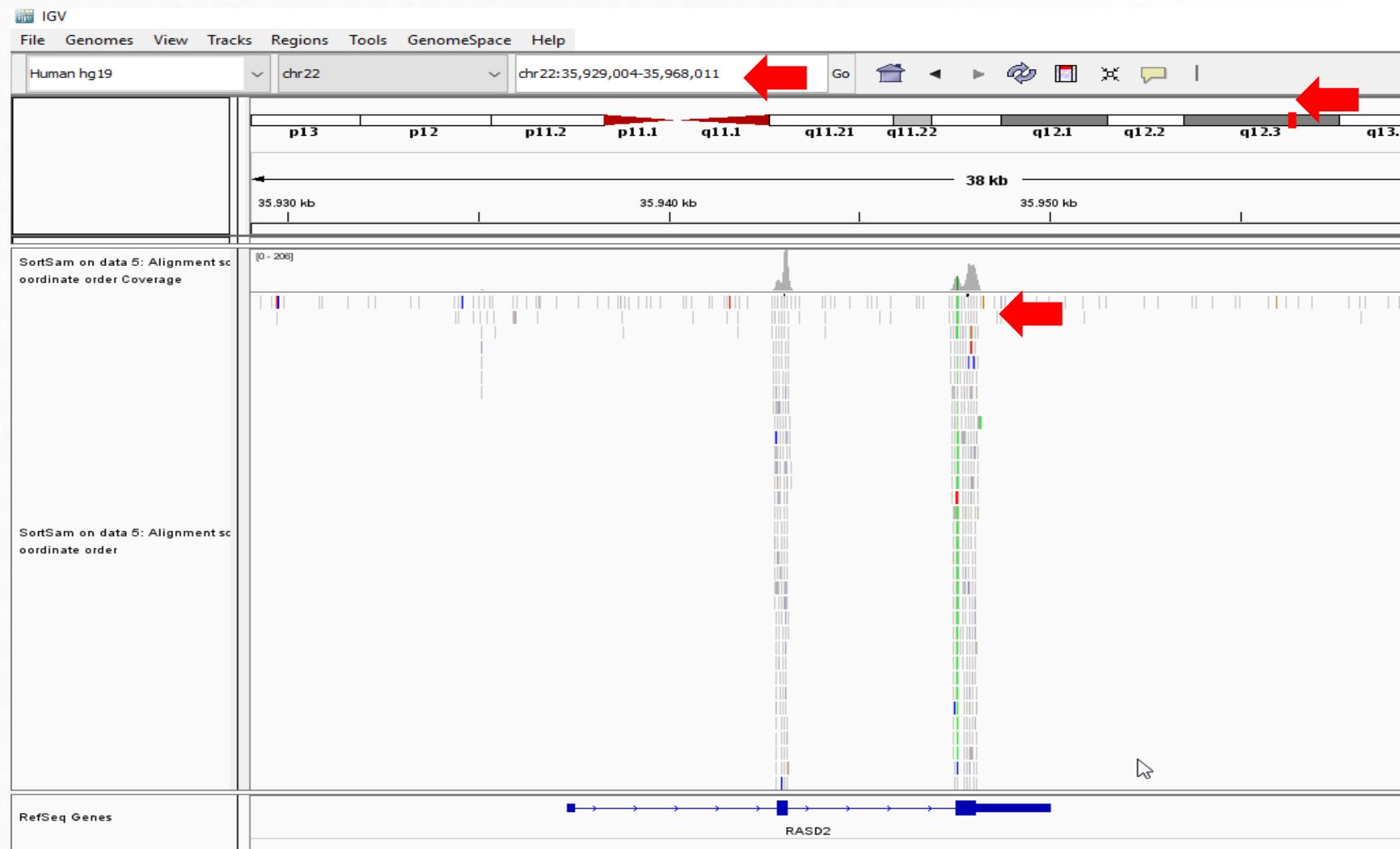


Command line IGV and igvtools for all platforms  
Separate Java 11 required

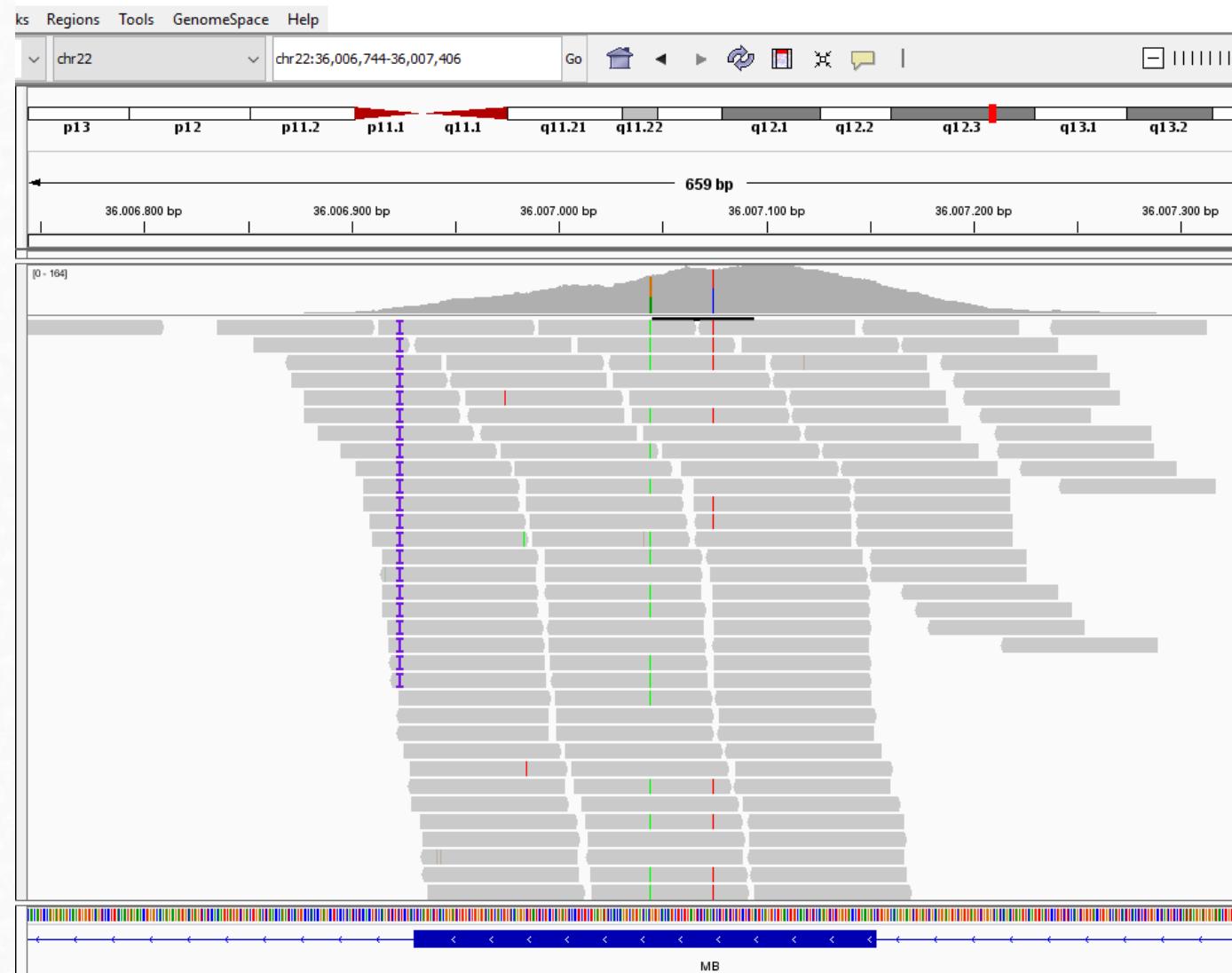


Load the .bam file

## 6. Visualize the BAM file with IGV



chr22:36,006,744-36,007,406



## 7. Generate a pileup file: a column-wise representation of the aligned reads

**Galaxy / Europe**

Tools

- [pileup](#)

[Get Data](#)

- [Download and Generate Pileup](#)
- [Format from NCBI SRA](#)

**SAM/BAM**

- [Filter pileup on coverage and SNPs](#)
- [Generate\\_pileup from BAM dataset](#)
- [samtools mpileup multi-way pileup of variants](#)
- [Pileup-to-Interval condenses pileup format into ranges of bases](#)
- [Convert\\_Merge\\_Randomize](#) BAM datasets and perform other transformations
- [Tag\\_pileup\\_frequency](#)

**VCF/BCF**

- [VarScan mpileup for variant detection](#)
- [VarScan copynumber](#) Determine relative tumor copy number from tumor-normal pileups
- [VarScan for variant detection](#)
- [bcftools call SNP/indel variant calling from VCF/BCF](#)

**Peak Calling**

- [MACS2 callpeak Call peaks from alignment results](#)

Analyze Data Workflow Visualize ▾ Shared Data ▾ Help ▾ User ▾

**Generate pileup from BAM dataset (Galaxy Version 1.1.2)**

Will you select a reference genome from your history or use a built-in index?

Use a built-in index

Select the BAM file to generate the pileup file for

6: SortSam on data 5: Alignment sorted in coordinate order  

Using reference genome

Human (Homo sapiens): hg19

Whether or not to print the mapping quality as the last column

Do not print the mapping quality as the last column

Makes the output easier to parse, but is space inefficient

Whether or not to print only output pileup lines containing indels

Print all lines

Where to cap mapping quality

60

Call consensus according to MAQ model?

Yes  

Theta parameter (error dependency coefficient) in the MAQ consensus calling model

0.85

Number of haplotypes in the sample

2

Greater than or equal to 2

Expected fraction of differences between a pair of haplotypes

0.001

Phred probability of an indel in sequencing/prep

40

Execute

summarises all data from the reads at each genomic region that is covered by at least one read. Each row of the pileup file gives similar information to a single vertical column of reads in the IGV view.

- Change file datatype

### Edit dataset attributes

Attributes updated.

≡ Attributes

⚙ Convert

🗄 Datatypes

👤 Permissions

#### Change datatype

➡ Change datatype

##### New Type

pileup

This will change the datatype of the existing dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

- Change file datatype

### Edit dataset attributes

Attributes updated.

☰ Attributes    ⚙ Convert    📁 Datatypes    🕸 Permissions

**Change datatype**    ↗ Change datatype

**New Type**

pileup

This will change the datatype of the existing dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.



Tip: The pileup file we generated has 10 columns:

1. chromosome
2. position
3. current reference base
4. consensus base from the mapped reads
5. consensus quality
6. SNV quality
7. maximum mapping quality
8. coverage
9. bases within reads
10. quality values

- . = match on forward strand for that base
- , = match on reverse strand
- ACGTN = mismatch on forward
- acgtm = mismatch on reverse
- +[0-9]+[ACGTNacgtm]+ = insertion between this reference position and the next
- -[0-9]+[ACGTNacgtm]+ = deletion between this reference position and the next
- ^ = start of read
- \$ = end of read
- BaseQualities = one character per base in ReadBases, ASCII encoded Phred scores

- Filter the pileup file

1	2	3	4	5	6	7	8	9
chr10	64870	c	C	30	0	0	1	^!,,
chr10	64871	t	T	30	0	0	1	,
chr10	64872	c	C	30	0	0	1	,
chr10	64873	a	A	30	0	0	1	,
chr10	64874	c	C	30	0	0	1	,
chr10	64875	t	T	30	0	0	1	,
chr10	64876	t	T	30	0	0	1	,
chr10	64877	t	T	30	0	0	1	,
chr10	64878	t	T	30	0	0	1	,
chr10	64879	t	T	30	0	0	1	,
chr10	64880	a	A	30	0	0	1	,
chr10	64881	a	A	30	0	0	1	,
chr10	64882	t	T	30	0	0	1	,
chr10	64883	a	A	30	0	0	1	,
chr10	64884	c	C	30	0	0	1	,
chr10	64885	t	T	30	0	0	1	,
chr10	64886	a	A	30	0	0	1	,
chr10	64887	a	A	30	0	0	1	,
chr10	64888	t	T	30	0	0	1	,
chr10	64889	g	G	30	0	0	1	,
chr10	64890	t	T	30	0	0	1	,
chr10	64891	t	T	30	0	0	1	,
chr10	64892	g	G	30	0	0	1	,
chr10	64893	g	G	30	0	0	1	,
chr10	64894	t	T	30	0	0	1	,
chr10	64895	t	T	30	0	0	1	,
chr10	64896	g	G	30	0	0	1	,
chr10	64897	t	T	30	0	0	1	,

- Filter Pileup file (filter out regions with coverage of at least 10 reads)

A lot of rows are from outside Chr 22 with very low coverage

**Galaxy / Europe**

Analyze Data Workflow Visualize ▾ Shared Data ▾ Help ▾ User ▾

**Tools**

**filter pileup**

**SAM/BAM**

**Filter pileup on coverage and SNPs**

Convert, Merge, Randomize BAM datasets and perform other transformations

**Peak Calling**

MACS2 callpeak Call peaks from alignment results

**Graph/Display Data**

JBrowse genome browser

**Workflows**

All workflows

**Filter pileup on coverage and SNPs (Galaxy Version 1.0.2)**

Select dataset: 10: Pileup File

which contains: Pileup with ten columns (with consensus) 

**Do not consider read bases with quality lower than:** 20

No variants with quality below this value will be reported

**Do not report positions with coverage lower than:** 10 

**Only report variants?** Yes

See "Examples 1 and 2" below for explanation

**Convert coordinates to intervals?** No

See "Output format" below for explanation

**Print total number of differences?** No

See "Example 3" below for explanation

**Print quality and base string?** Yes

See "Example 4" below for explanation

**Execute**

chr21	15441242	A	A	187	0	19	62	.....T,.....
chr21	15441261	T	T	32	0	5	16	CC,,C,cCcc,,CCC
chr21	15441264	T	T	11	0	5	16	CCCCCcccCCCCCCCC
chr21	48110986	C	C	53	0	0	16	,.....g
chr22	16084443	G	G	56	0	0	18	.....a
chr22	16084718	G	G	27	0	0	10	A.A.AA..AA
chr22	16084749	G	G	41	0	0	23	..T.....^!,^!,^!,
chr22	16084859	G	G	37	0	1	12	..,a,,a,,a,,
chr22	16100462	A	A	38	0	0	18	....G...GG....
chr22	16100514	C	C	46	0	0	39	.....,g,,,G,.....,g,...^!,
chr22	16100525	T	T	61	0	0	45	....C,.....,g^!,
chr22	16123048	G	G	54	0	0	78	.\$.....,a.....,
chr22	16123080	G	G	70	0	0	73	.....,C.....,
chr22	16123110	C	C	82	0	0	52	.\$,\$,\$,,A,,
chr22	16123164	C	C	53	0	0	28	a,,a,A,,A,,
chr22	16123175	C	C	48	0	0	27	.....,g,g,...^!,
chr22	16123177	T	T	50	0	0	26	,\$,,C.....,
chr22	16123245	T	T	68	0	6	106	.\$.....,g..gg,...g,..{
chr22	16123252	T	T	97	0	7	110	.GG,G.,CGG,,G.G.G.,G.NGG,,G.G.GG,,G.....G
chr22	16123266	C	C	112	0	9	134	.....,t,,G,.....,T,.....g,,
chr22	16123289	G	G	121	0	10	144	.....,
chr22	16123296	T	T	109	0	10	141	C\$,\$,,c,,c,,c,,cC,,c,,c,,c,c,,
chr22	16123469	T	T	63	0	19	16	.\$g,.....,
chr22	16123495	C	C	80	0	16	22	,\$,.....,a,,
chr22	16123524	A	A	64	0	5	20	.....,C,C,,
chr22	16123608	A	A	57	0	0	39	,\$,,t,.....,^!,
chr22	16123629	G	G	58	0	0	42	T,.....,t,,T,,
chr22	16123722	A	A	194	0	15	183	.....,G,,
chr22	16123732	C	C	230	0	14	225	,\$,.....,
chr22	16123743	G	G	235	0	13	246	.....,
chr22	16123766	T	T	231	0	15	194	.\$,\$,\$,\$,\$,.....,C,,
chr22	16123780	C	C	227	0	16	147	.\$,\$,\$,,.....,t,,
chr22	16123793	G	G	96	0	9	100	.\$,,C,,A,,aa,,A,,a,a,a,,aa,,AA,,aa,A,,A,,a,,

## 8. Call variants with FreeBayes

**Galaxy / Europe**

Analyze Data Workflow Visualize Shared Data Help User Versions Options

**Tools**

freebayes

**VCF/BCF**

FreeBayes bayesian genetic variant detector

BamLeftAlign indels in BAM datasets

BamLeftAlign indels in BAM datasets

**Mapping**

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

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

**Variant Calling**

BamLeftAlign indels in BAM datasets

FreeBayes bayesian genetic variant detector

**SNPEFF**

SnpEff build: database from Genbank or GFF record

**FreeBayes bayesian genetic variant detector (Galaxy Version 1.1.0.46-0)**

Choose the source for the reference genome

Locally cached

Run in batch mode?

Run individually

Merge output VCFs

Selecting individual mode will generate one VCF dataset for each input BAM dataset. Selecting the merge option will produce one VCF dataset for all input BAM datasets

**BAM dataset**

6: SortSam on data 5: Alignment sorted in coordinate order

Using reference genome

Human (Homo sapiens): hg19

Limit variant calling to a set of regions?

Do not limit

Sets --targets or --region options

Choose parameter selection level

2. Simple diploid calling with filtering and coverage

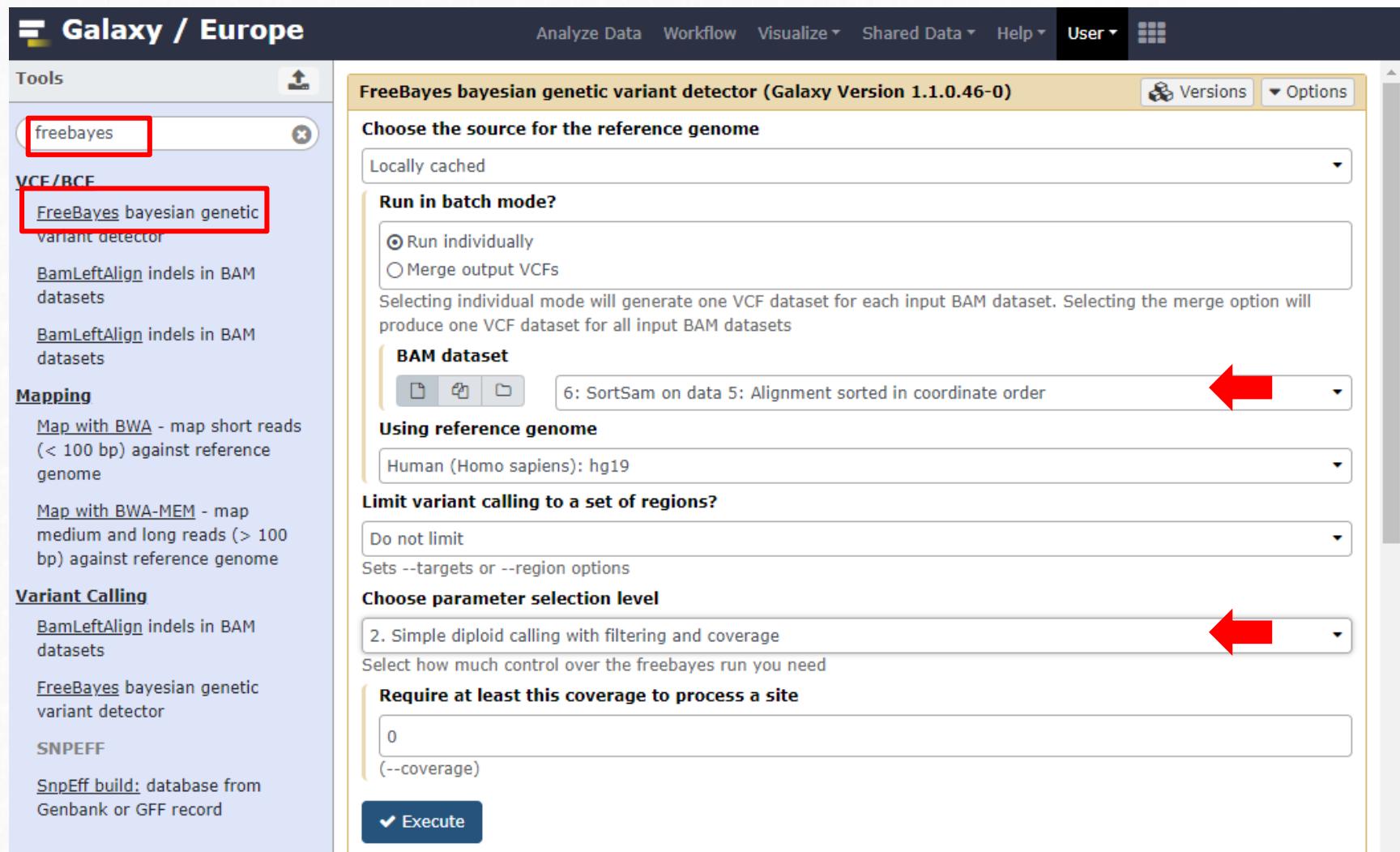
Select how much control over the freebayes run you need

Require at least this coverage to process a site

0

(--coverage)

Execute



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr11	116691511.		GGACAGACAGACAGACAG	GGACAGACAGACAG	3109.94	.	AB
chr11	116691634.	C		A	267384.	.	AB
chr11	116692334.	C		T	744501.	.	AB
chr11	116692694.	G		A	6855.03	.	AB
chr11	116693464.	C		T	5143.08	.	AB
chr11	116697848.	G		A	41.8186	.	AB
chr11	116701535.	T		C	1782.38	.	AB
chr11	116703640.	G		C	1936.77	.	AB
chr11	116703671.	G		T	1035.97	.	AB
chr11	116707583.	A		G	192112.	.	AB
chr11	116707684.	A		G	2180.07	.	AB
chr11	116708020.	A		G	50691.	.	AB
chr11	116720089.	T		C	354637.	.	AB
chr11	116720137.	G		C	75.4953	.	AB
chr16	32486478.	T		C	49.4196	.	AB
chr16	32486483.	CATC		TATT	46.8065	.	AB
chr16	32486492.	A		T	67.9303	.	AB
chr16	32486517.	C		T	74.8098	.	AB
chr16	32486534.	C		T	45.3812	.	AB
chr18	14183638.	G		C	206811.	.	AB
chr19	9060294.	G		T	19.9839	.	AB
chr2	95513809.	C		T	109935.	.	AB
chr2	95513817.	G		T	121382.	.	AB
chr2	132367062.	G	TTTTTTTTTG	GTTTTTTTTG	66.9106	.	AB
chr22	16350323.	T		C	5.21218	.	AB
chr22	16350349.	G		C	49.2459	.	AB
chr22	16868364.	G		A	50.9845	.	AB
chr22	16870890.	C		T	91271.	.	AB
chr22	16871440.	A		C	44.2351	.	AB
chr22	17054103.	G		A	10109.	.	AB
chr22	17055569.	T		G	204408.	.	AB
chr22	17076273.	G		A	448374.	.	AB
chr22	17119450.	G		A	0.0135665	.	AB
chr22	17127617.	A		G	67.1813	.	AB
chr22	17309881.	A		G	549584.	.	AB
chr22	17326668.	A		G	125863.	.	AB
chr22	17339003.	G		A	420541.	.	AB
chr22	17339041.	G		A	39.6615	.	AB
chr22	17339068.	T		C	97.3915	.	AB
chr22	17339129.	C		G	504.11	.	AB

Col	Field	Description
1	CHROM	Chromosome name
2	POS	1-based position. For an indel, this is the position preceding the indel.
3	ID	Variant identifier (optional). Usually the dbSNP rsID.
4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
5	ALT	Comma delimited list of alternative sequence(s) seen in our reads.
6	QUAL	Phred-scaled probability of all samples being homozygous reference.
7	FILTER	Semicolon delimited list of filters that the variant fails to pass.
8	INFO	Semicolon delimited list of variant information.
9	FORMAT	Colon delimited list of the format of individual genotypes in the following fields.
10+	Sample(s)	Individual genotype information defined by FORMAT.

- Visualise the VCF file with IGV

[12: FreeBayes on data 6 \(variants\)](#)   

2,455 lines, 153 comments  
format: **vcf**, database: **hg19**

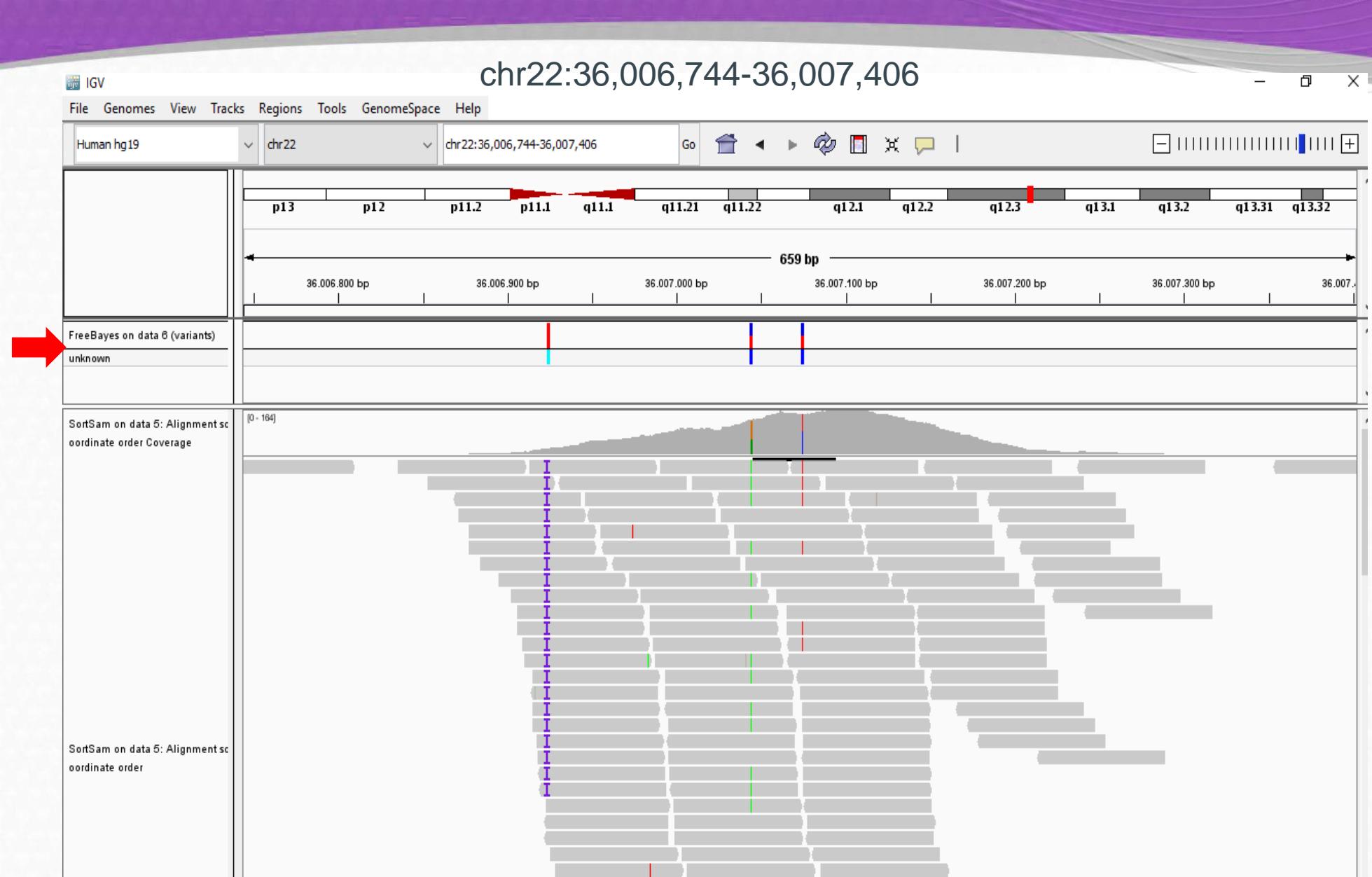
      

[play at UCSC main](#)  
[play with IGV local Human hg19](#)  
[display at RViewer main](#)

**1.Chrom**

```
##fileformat=VCFv4.2
##fileDate=20190220
##source=FreeBayes v1.1.0-46-g8d2b3a0-dirty
##reference=/data/db/reference_genomes/
##contig=<ID=chr10,length=135534747>
```

<  >



- Filter VCF file

**Galaxy Europe**

Analyze Data Workflow Visualize ▾ Shared Data ▾ Help ▾ User ▾  

**Tools**

**filter** 

Show Sections

provided with rxDock

**Filter segmentation** Filter segmentation by rules

**Filter FASTQ** reads by quality score and length

**Stacks: clone filter** Identify PCR clones

**Stacks2: clone filter** Identify PCR clones

**SnpSift** Filter variants using arbitrary expressions

**Filter GTF data by attribute values\_list**

**Aggregate and filter alignment metrics** of individual clusters, like the output of graphclust\_align\_cluster

**FilterSamReads** include or exclude aligned and unaligned reads and read lists

**Kraken-filter** filter classification by confidence score

**Filter** data on any column using simple expressions 

Filter data on any column using simple expressions (Galaxy Version 1.1.1)

Favorite Options

Dataset missing? See TIP below.

**With following condition**

c1=='chr22' and c2 > 36006744 and c2 < 36007406 

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

**Number of header lines to skip**

0

**Email notification**

Yes No

Send an email notification when the job completes.

Execute

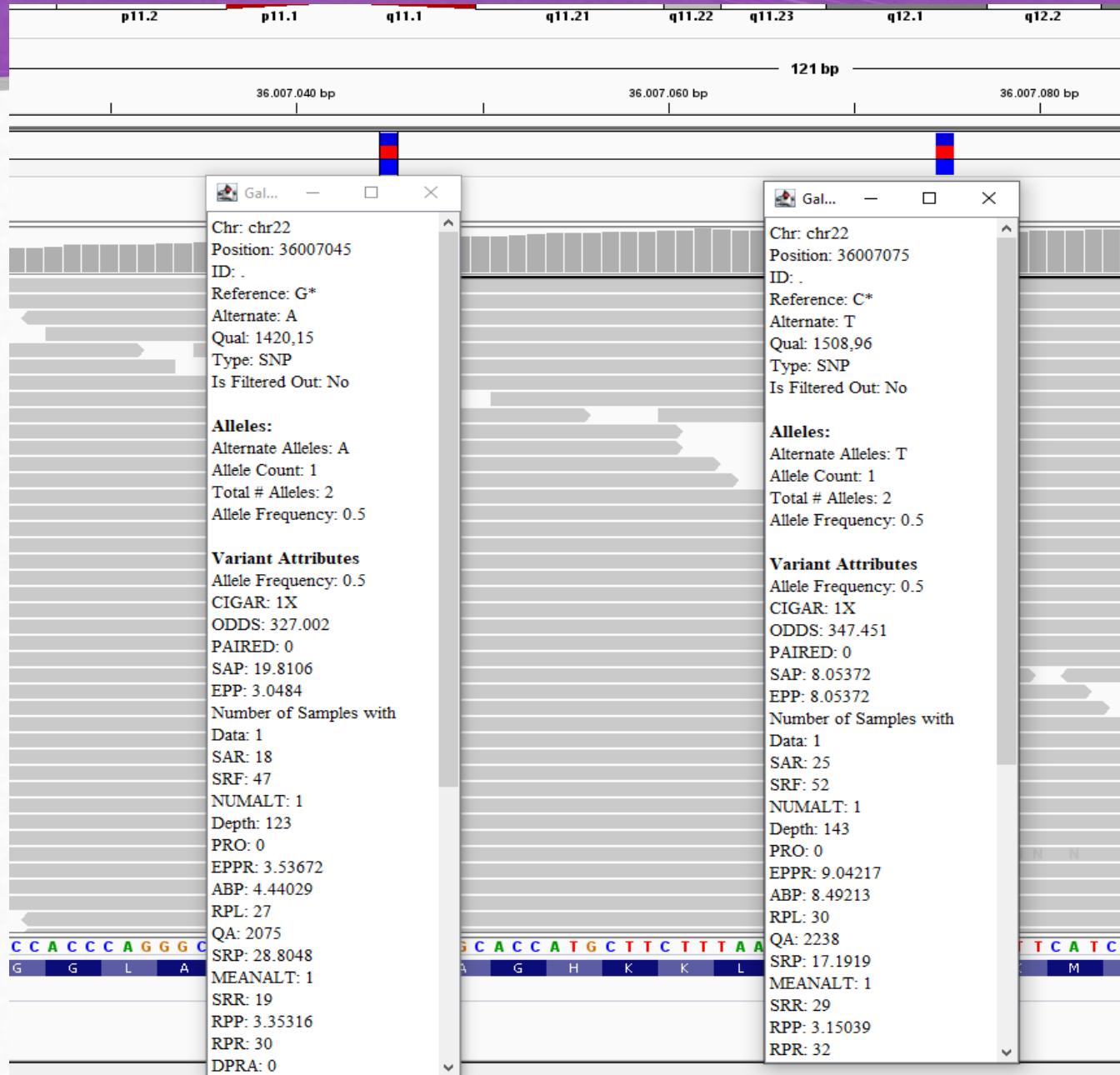
**⚠ Double equal signs. ==, must be used as "equal to" (e.g., c1 == 'chr22')**

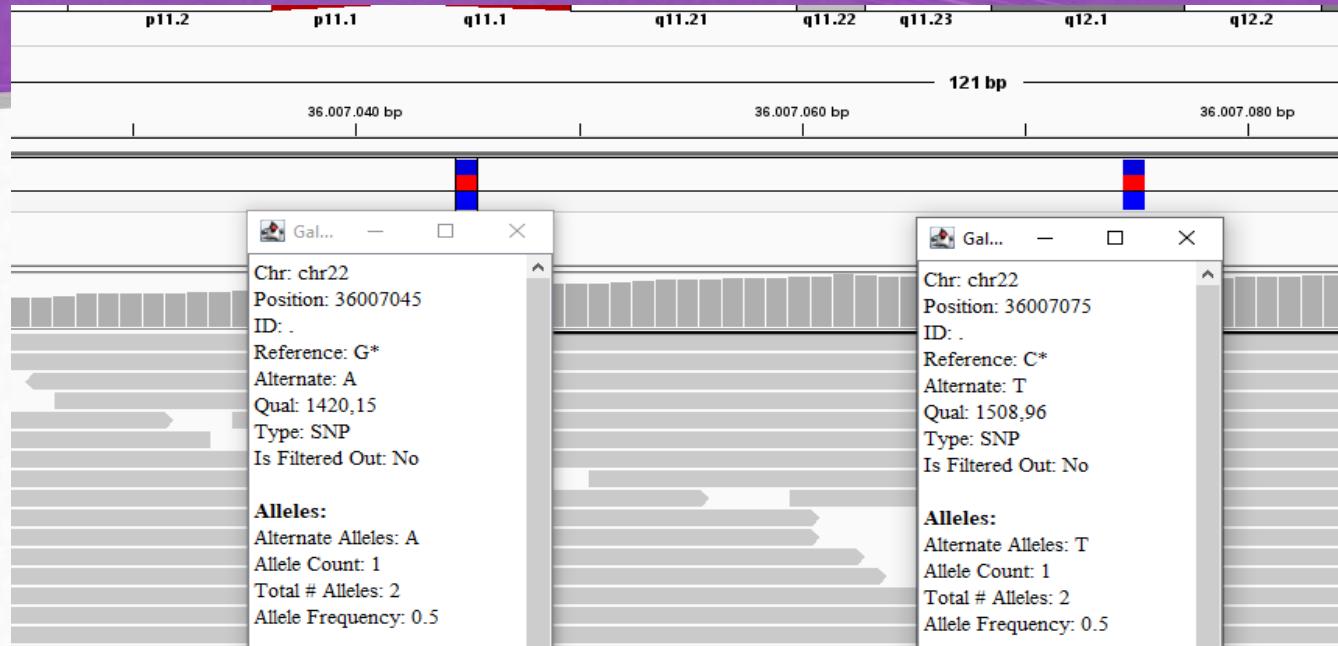
**ℹ TIP:** Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

**ℹ TIP:** If your data is not TAB delimited, use Text Manipulation->Convert

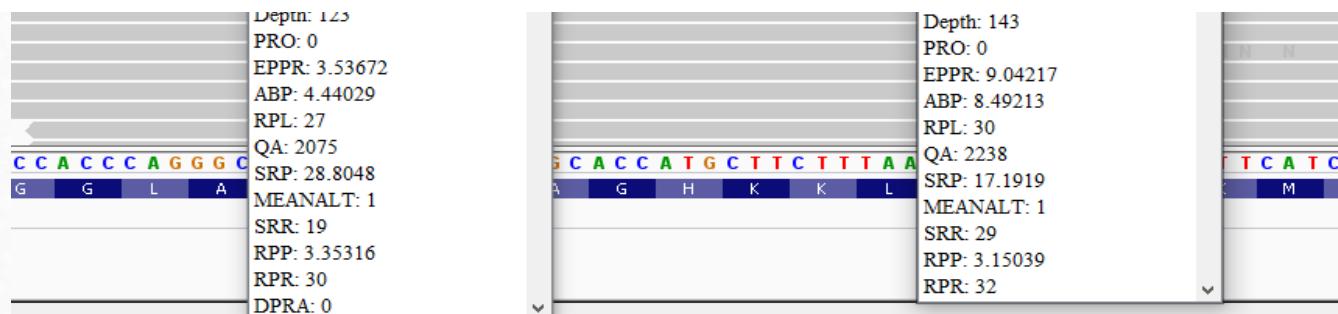
c1=='chr22' and c2 > 36006744 and c2 < 36007406

Chrom	Pos	ID	Ref	Alt	Qual	Filter	Info
chr22	36006923	.	GCT	GCCT	645.607	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=21;CIGAR=1M12M;DP=21;DPB=28;DPRA=0;EPP=5.59539;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;
chr22	36007045	.	G	A	1420.15	.	AB=0.463415;ABP=4.44029;AC=1;AF=0.5;AN=2;AO=57;CIGAR=1X;DP=123;DPB=123;DPRA=0;EPP=3.0484;EPPR=3.53672;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;
chr22	36007075	.	C	T	1508.96	.	AB=0.433566;ABP=8.49213;AC=1;AF=0.5;AN=2;AO=62;CIGAR=1X;DP=143;DPB=143;DPRA=0;EPP=8.05372;EPPR=9.04217;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;





Chrom	Pos	ID	Ref	Alt	Qual	Filter	Info
chr22	36006923	.	GCT	GCCT	645.607	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=21;CIGAR=1M1I2M;DP=21;DPB=28;DPRA=0;EPP=5.59539;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;N
chr22	36007045	.	G	A	1420.15	.	AB=0.463415;ABP=4.44029;AC=1;AF=0.5;AN=2;AO=57;CIGAR=1X;DP=123;DPB=123;DPRA=0;EPP=3.0484;EPPR=3.53672;GTI=0;LEN=1;MEANALT=1;MQ
chr22	36007075	.	C	T	1508.96	.	AB=0.433566;ABP=8.49213;AC=1;AF=0.5;AN=2;AO=62;CIGAR=1X;DP=143;DPB=143;DPRA=0;EPP=8.05372;EPPR=9.04217;GTI=0;LEN=1;MEANALT=1;MQ



## 9. Annotating variants with SnpEff

Genetic variant annotation and functional effect prediction toolbox. It annotates and predicts the effects of genetic variants on genes and proteins (such as amino acid changes).

<http://snpeff.sourceforge.net/>

**Typical usage :**

- **Input:** The inputs are predicted variants (SNPs, insertions, deletions and MNPs). The input file is usually obtained as a result of a sequencing experiment, and it is usually in variant call format (VCF).
- **Output:** SnpEff analyzes the input variants. It annotates the variants and calculates the effects they produce on known genes (e.g. amino acid changes).

## 9. Annotating variants with SnpEff

**Galaxy / Europe**

Analyze Data Workflow Visualize Shared Data Help User Versions Options

**Tools**

- snpeff** (highlighted with a red box)
- VCF/BCF**
  - snippy**: Snippy finds SNPs between a haploid reference genome and your NGS sequence reads.
- Annotation**
  - SnpSift GeneSets**: Annotating GeneSets, such as Gene Ontology, KEGG, Reactome
- Variant Calling**
  - SNPEFF**
    - SnpEff eff: annotate variants** (highlighted with a red box)
    - SnpEff build**: database from Genbank or GFF record
    - SnpEff databases**: list available databases
    - SnpEff download**: download a pre-built database
    - SnpEff Ensembl CDS Report**: Variant coding sequence changes for SnpEffects
    - SnpEff to Peptide fasta**: to create a Search DB fasta for variant SAP peptides

**SnpEff eff: annotate variants (Galaxy Version 4.3+T.galaxy1)**

Sequence changes (SNPs, MNPs, InDels)  
 12: FreeBayes on data 6 (variants) (highlighted with a red arrow)

**Input format**  
 VCF (highlighted with a red arrow)

**Output format**  
 VCF (only if input is VCF)

**Create CSV report, useful for downstream analysis (-csvStats)**  
 Yes No

**Genome source**  
 Locally installed SnpEff database

**Genome**  
 Homo sapiens : hg19 (highlighted with a red arrow)

**Regulation options**

**Upstream / Downstream length**  
 5000 bases  
 (-ud)

**Set size for splice sites (donor and acceptor) in bases**  
 2 bases  
 (-ss)

**spliceRegion Settings**  
 Use Defaults

**Annotation options**

# 1. Hands-On Exome Variant Analysis

SnpEff will generate two outputs:

- an annotated VCF file
- an HTML report

## SnpEff: Variant analysis



0|,A|intron\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|14/20|c.3154-29C>T||||||  
3|||||2249|,G|intron\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|13/20|c.3067-112T>C|||  
T|POLRMT|transcript|NM\_005035.3|protein\_coding|13/20|c.3066+12A>C||||||  
.0|c.2887-7C>G|||||,C|downstream\_gene\_variant|MODIFIER|HCN2|HCN2|transcript|NM\_001194.3|protein\_coding||c.\*329  
>C|||||2754|,C|intron\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|12/20|c.2886+45A>G|  
c.2840A>G|p.Glu947Gly|2896/3800|2840/3693|947/1230||,C|downstream\_gene\_variant|MODIFIER|HCN2|HCN2|transcript  
T>A|p.Ala933Ala|2855/3800|2799/3693|933/1230||,T|downstream\_gene\_variant|MODIFIER|HCN2|HCN2|transcript|NM\_00  
?7A>C|||||3042|,C|intron\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|11/20|c.2764-121T  
\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|11/20|c.2764-130T>G||||||  
.3740T>C|||||3055|,C|intron\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|11/20|c.2764-  
c.\*3754A>C|||||3069|,C|intron\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|11/20|c.2763  
|||||3140|,A|intron\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|11/20|c.2763+66G>T||||||  
|||3156|,C|intron\_variant|MODIFIER|POLRMT|POLRMT|transcript|NM\_005035.3|protein\_coding|11/20|c.2763+50T>G||||||  
n\_coding|11/21|c.2747A>C|||||,G|structural\_interaction\_variant|HIGH|POLRMT|POLRMT|interaction|4BOCA\_827-A\_916:  
70/3800|2714/3693|905/1230||,C|downstream\_gene\_variant|MODIFIER|HCN2|HCN2|transcript|NM\_001194.3|protein\_codi  
/3800|2699/3693|900/1230||,C|downstream\_gene\_variant|MODIFIER|HCN2|HCN2|transcript|NM\_001194.3|protein\_coding  
|5.3|protein\_coding|11/21|c.2674T>G|||||,C|structural\_interaction\_variant|HIGH|POLRMT|POLRMT|interaction|3SPA:A\_81|  
59A>G|p.Glu890Gly|2725/3800|2669/3693|890/1230||,C|downstream\_gene\_variant|MODIFIER|HCN2|HCN2|transcript|NM\_

## Summary

Contents	
<a href="#">Summary</a>	
<a href="#">Variant rate by chromosome</a>	
<a href="#">Variants by type</a>	
<a href="#">Number of variants by impact</a>	
<a href="#">Number of variants by functional class</a>	
<a href="#">Number of variants by effect</a>	
<a href="#">Quality histogram</a>	
<a href="#">InDel length histogram</a>	
<a href="#">Base variant table</a>	
<a href="#">Transition vs transversions (ts/tv)</a>	
<a href="#">Allele frequency</a>	
<a href="#">Allele Count</a>	
<a href="#">Codon change table</a>	
<a href="#">Amino acid change table</a>	
<a href="#">Chromosome variants plots</a>	
<a href="#">Details by gene</a>	

Genome	hg19
Date	2020-12-01 21:42
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /data/dnb03/galaxy_db/job_working_directory/013/520/13520986/outputs/galaxy.
Warnings	40
Errors	0
Number of lines (input file)	2,583
Number of variants (before filter)	2,590
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	2,590
Number of known variants (i.e. non-empty ID)	0 ( 0 % )
Number of multi-allelic VCF entries (i.e. more than two alleles)	7
Number of effects	7,255
Genome total length	3,137,161,265
Genome effective length	657,375,723
Variant rate	1 variant every 253,813 bases

Genome	hg19
Date	2020-12-01 21:42
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /data/dnb03/galaxy_db/job_working_directory/013/520/13520986/outputs/galaxy
Warnings	40
Errors	0
Number of lines (input file)	2,583
Number of variants (before filter)	2,590
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	2,590
Number of known variants (i.e. non-empty ID)	0 ( 0% )
Number of multi-allelic VCF entries (i.e. more than two alleles)	7
Number of effects	7,255
Genome total length	3,137,161,265
Genome effective length	657,375,723
Variant rate	1 variant every 253,813 bases

#### Number variants by type

Type	Total
SNP	2,389
MNP	61
INS	60
DEL	73
MIXED	7
INV	0
DUP	0
BND	0
INTERVAL	0
Total	2,590

Genome	hg19
Date	2020-12-01 21:42
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /data/dnb03/galaxy_db/job_working_directory/013/520/13520986/outputs/galaxy
Warnings	40
Errors	0
Number of lines (input file)	2,583
Number of variants (before filter)	2,590
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	2,590
Number of known variants (i.e. non-empty ID)	0 ( 0 % )

#### Variants rate details

Chromosome	Length	Variants	Variants rate
2	243,199,373	1	243,199,373
11	135,006,516	16	8,437,907
16	90,354,753	5	18,070,950
18	78,077,248	1	78,077,248
19	59,128,983	1	59,128,983
22	51,304,566	2,521	20,350
Un_gl000211	166,566	38	4,383
Un_gl000214	137,718	7	19,674
Total	657,375,723	2,590	253,813

#### Number variants by type

Type	Total
SNP	2,389
MNP	61
INS	60
DEL	73
MIXED	7
INV	0
DUP	0
BND	0
INTERVAL	0
Total	2,590

#### Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	276	3.804%
LOW	647	8.918%
MODERATE	565	7.788%
MODIFIER	5,767	79.49%

#### Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	500	51.867%
NONSENSE	1	0.104%
SILENT	463	48.029%