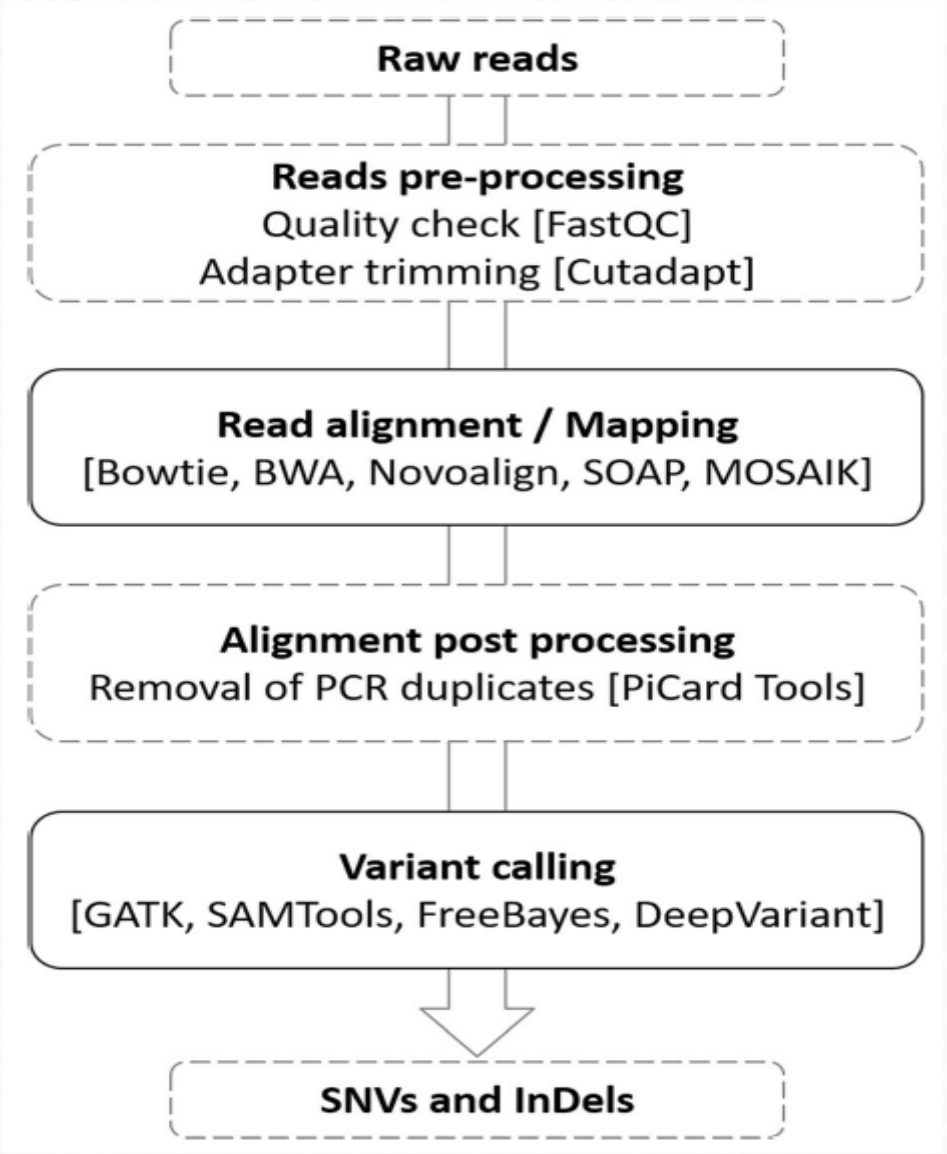
Vall d'Hebron
Institut de Recerca

HANDS ON: INTRODUCTION TO VARIANT ANALYSIS

Bioinformàtica per a la Recerca Biomèdica

Mireia Ferrer[1], Álex Sánchez[1,2]
Esther Camacho[1], Angel Blanco[1,2]
1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR
2 Departament de Genètica, Microbiologia i Estadística, UB

Vall d'Hebron Hospital · Generalitat de Catalunya Departament de Salut · Institut Català de la Salut · UAB Universitat Autònoma de Barcelona · Generalitat de Catalunya Departament d'Economia i Coneixement · Institució CERCA Centres de Recerca de Catalunya · Instituto de Salud Carlos III



Kumaran *et al. BMC Bioinformatics* (2019) 20:342
https://doi.org/10.1186/s12859-019-2928-9

This presentation is mainly based in a tutorial published by:

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/

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

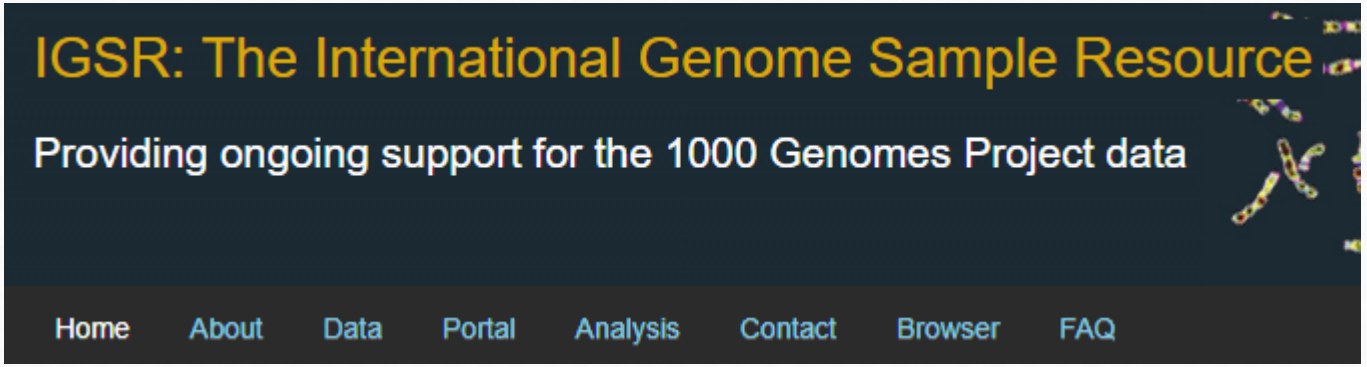
Vall d'Hebron
Institut de Recerca

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



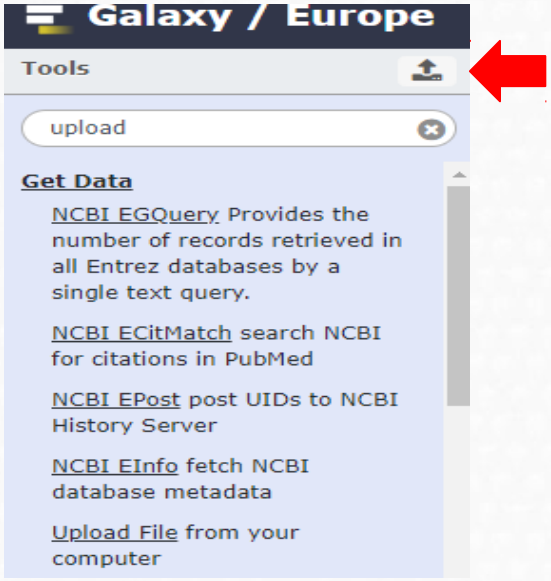
1. Open Galaxy and create a new history and name it (“Exon Variant Analysis”)
   https://usegalaxy.eu/

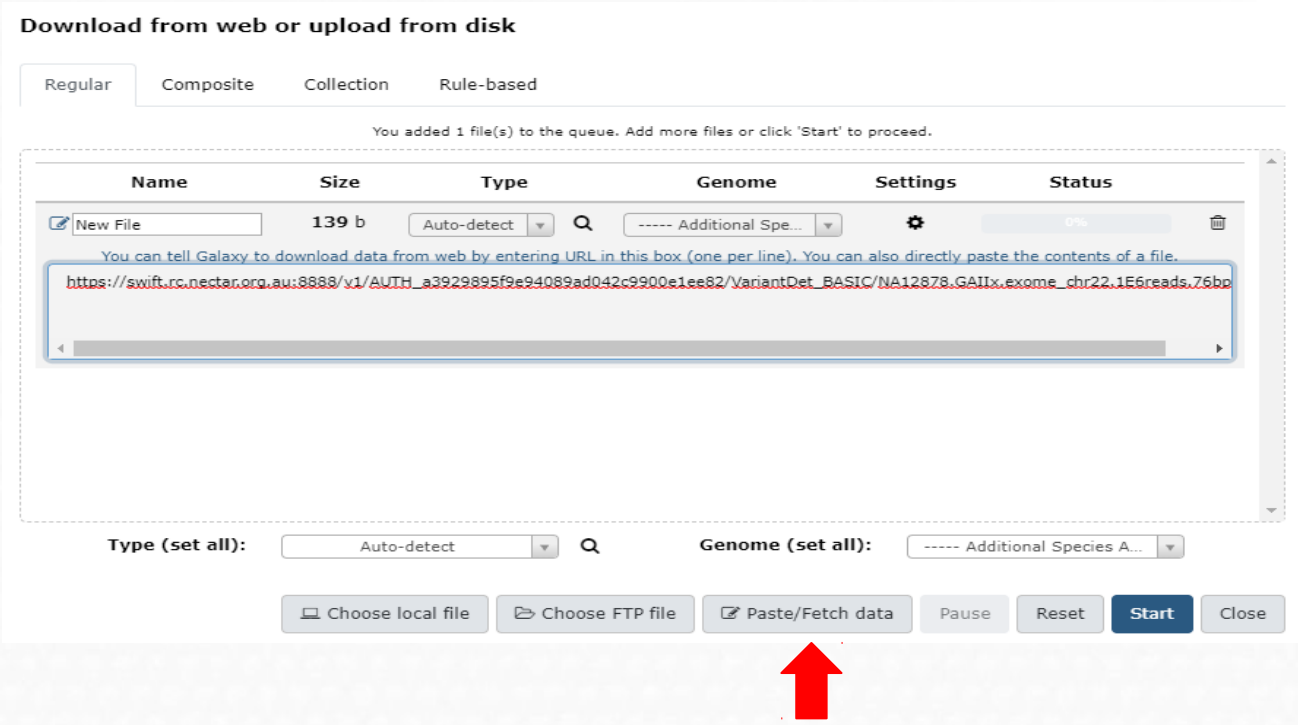
   https://usegalaxy.eu/join-training/ueb_bi2022/

Import data for the tutorial

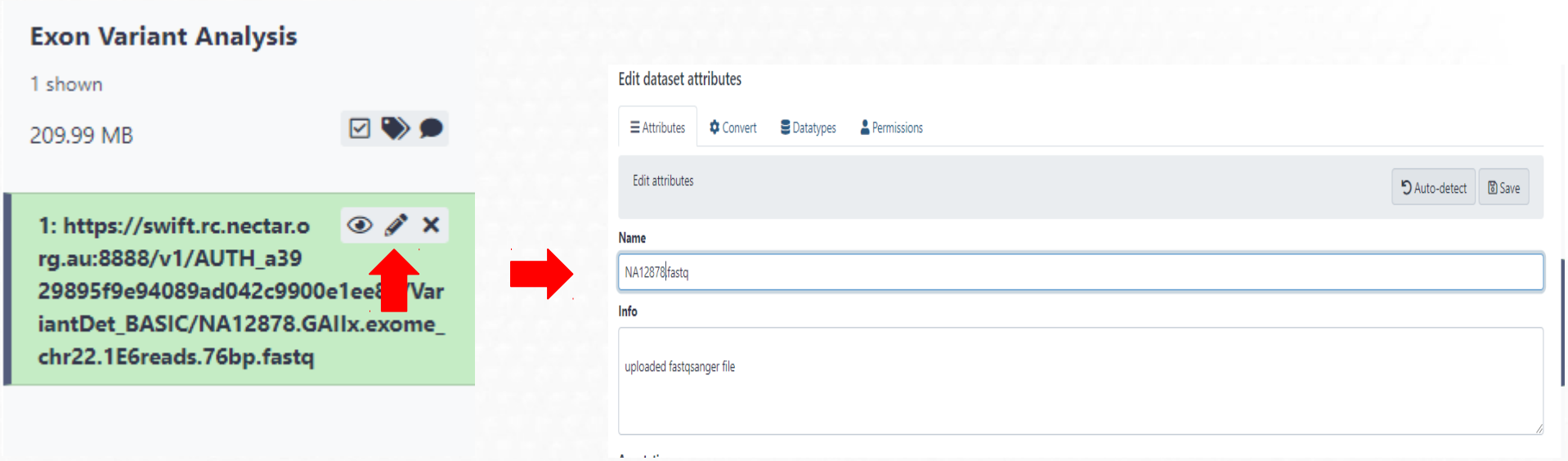
1. Paste the following link to download the data

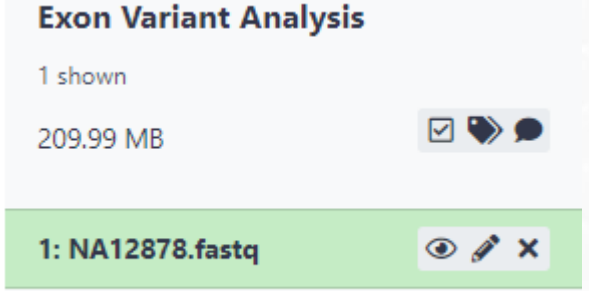
https://swift.rc.nectar.org.au:8888/v1/AUTH_a3929895f9e94089ad042c9900e1ee82/VariantDet_BASIC/NA12878.GAIIx.exome_chr22.1E6reads.76bp.fastq





Change the name of file:

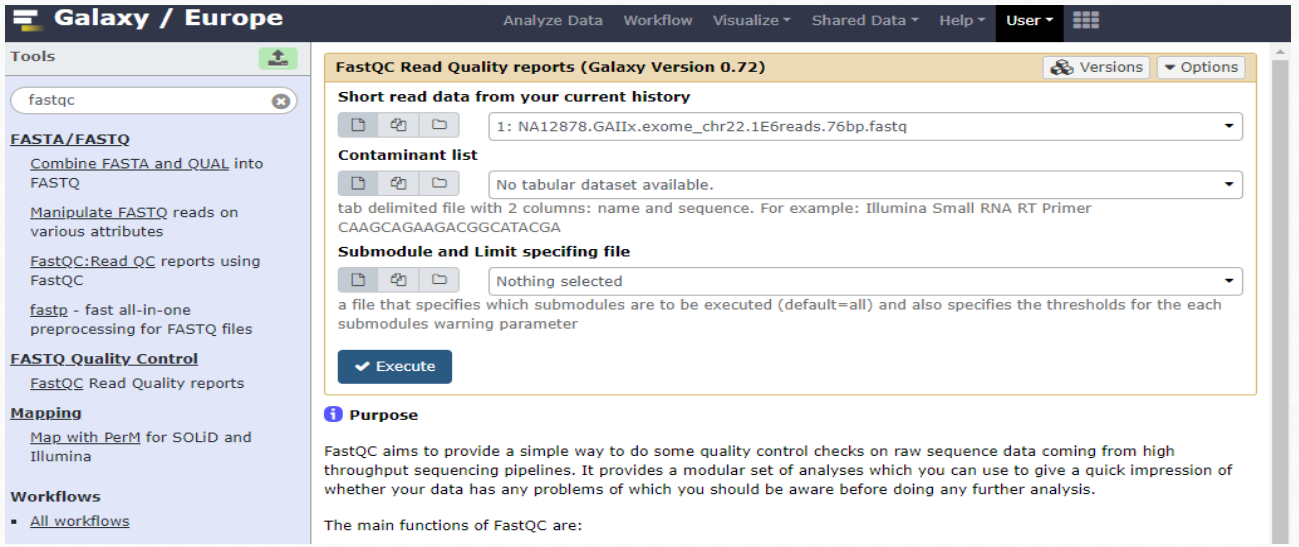




- Take a look at the FASTQ file

```
@61CC3AAXX100125:7:118:2538:5577/1
GACACCTTTAATGTCTGAAAAGAGACATTCACCATCTATTCTCTTGGAGGGCTACCACCTAAGAGCCTTCATCCCC
+
?>CADFEEEBEDIEHHIDGGGEEEEHFFGIGIIFFIIEFHIIIIHIIFFIIIDEIIGIIIEHFFFIIEHIFA@?==
```

3. Assessing read quality from the FASTQ files

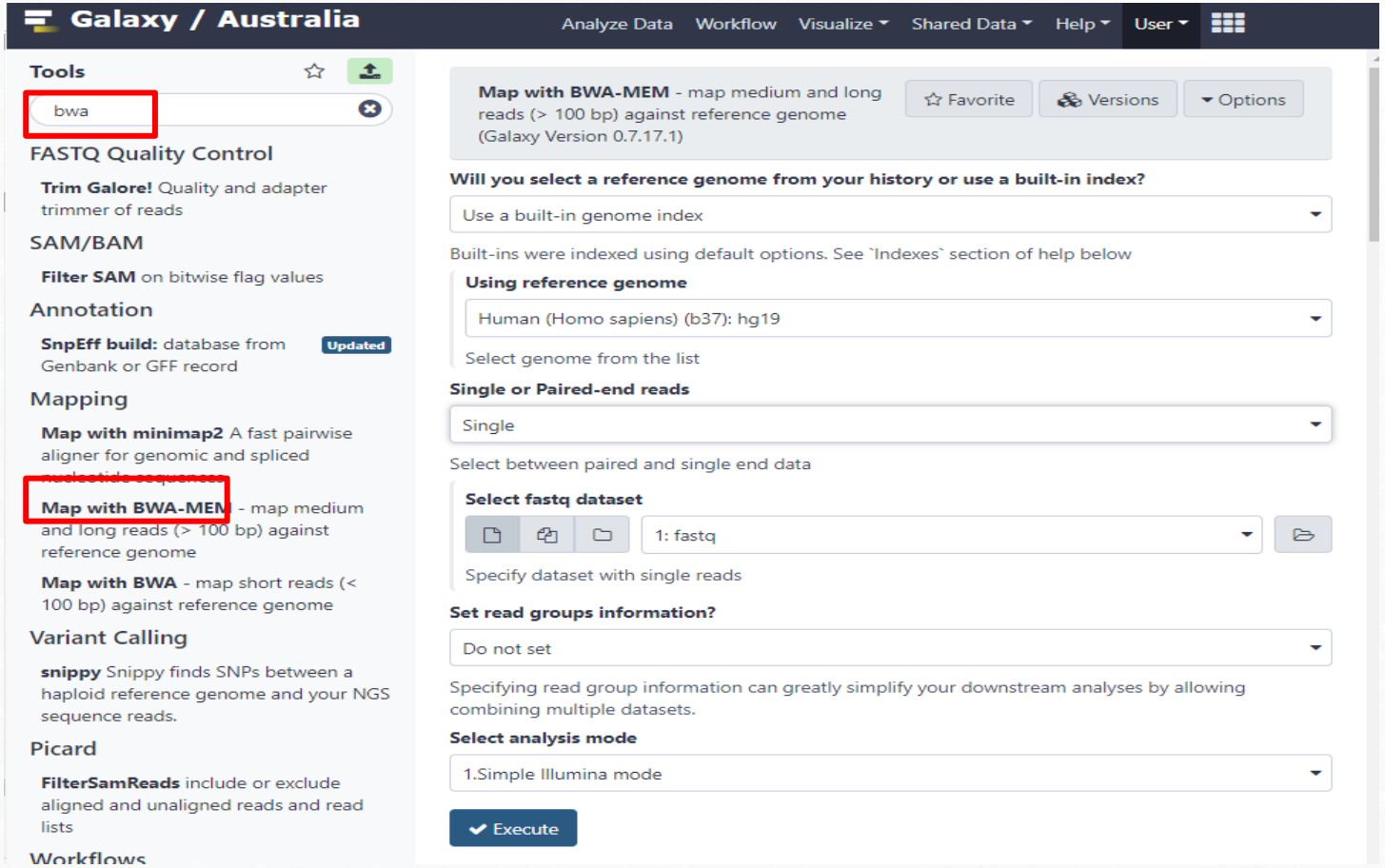


- Results of the Quality Control

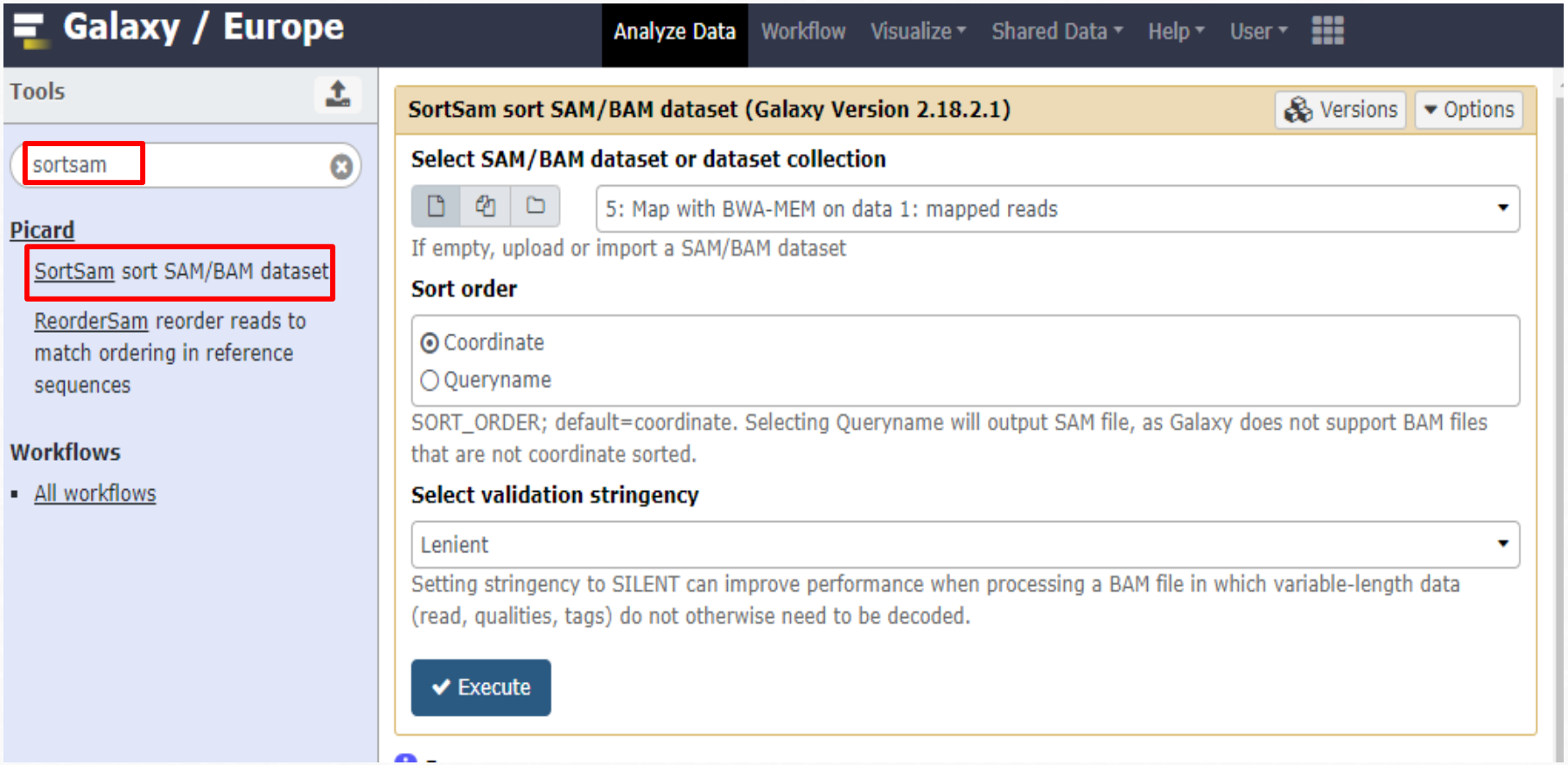
OR

4. Align the reads with BWA

Reference Genome: *Human genome 19 (hg19)*



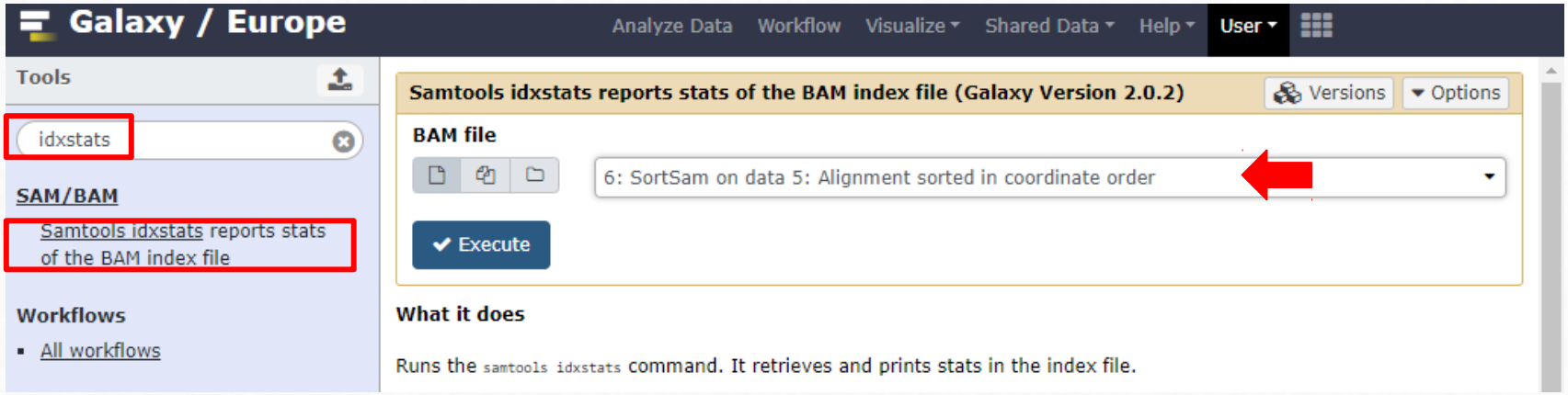
- Sort the BAM file



61CC3AAXX100125:7:118:2538:5577	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



generate some mapping statistics from the BAM file

ion ------ ---------- 1 Reference sequence identifier 2 Reference sequence length 3 Number of mapped reads 4 Number of placed but unmapped reads (typically unmapped partners of

Vall d'Hebron
Institut de Recerca

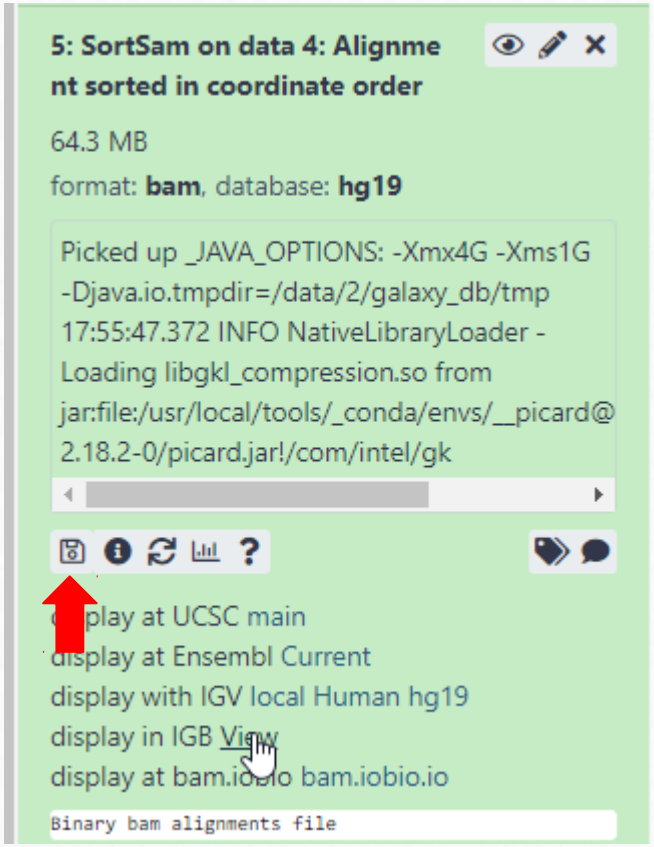
1	2	3	4
chr10	135534747	786	0
chr11	135006516	5541	0
chr11_gl000202_random	40103	0	0
chr12	133851895	534	0
chr13	115169878	199	0
chr14	107349540	12913	0
chr15	102531392	621	0
chr16	90354753	724	0
chr17_ctg5_hap1	1680828	2	0
chr17	81195210	275	0
chr17_gl000203_random	37498	0	0
chr17_gl000204_random	81310	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	[illegible]	0
chr22	51304566	1101584	0
chr2	243199373	[illegible]	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

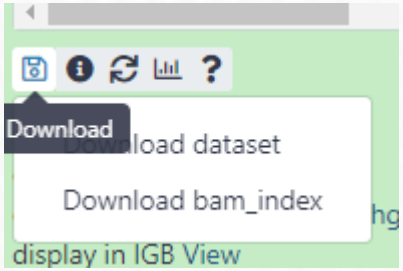
Column Description

------ ----------

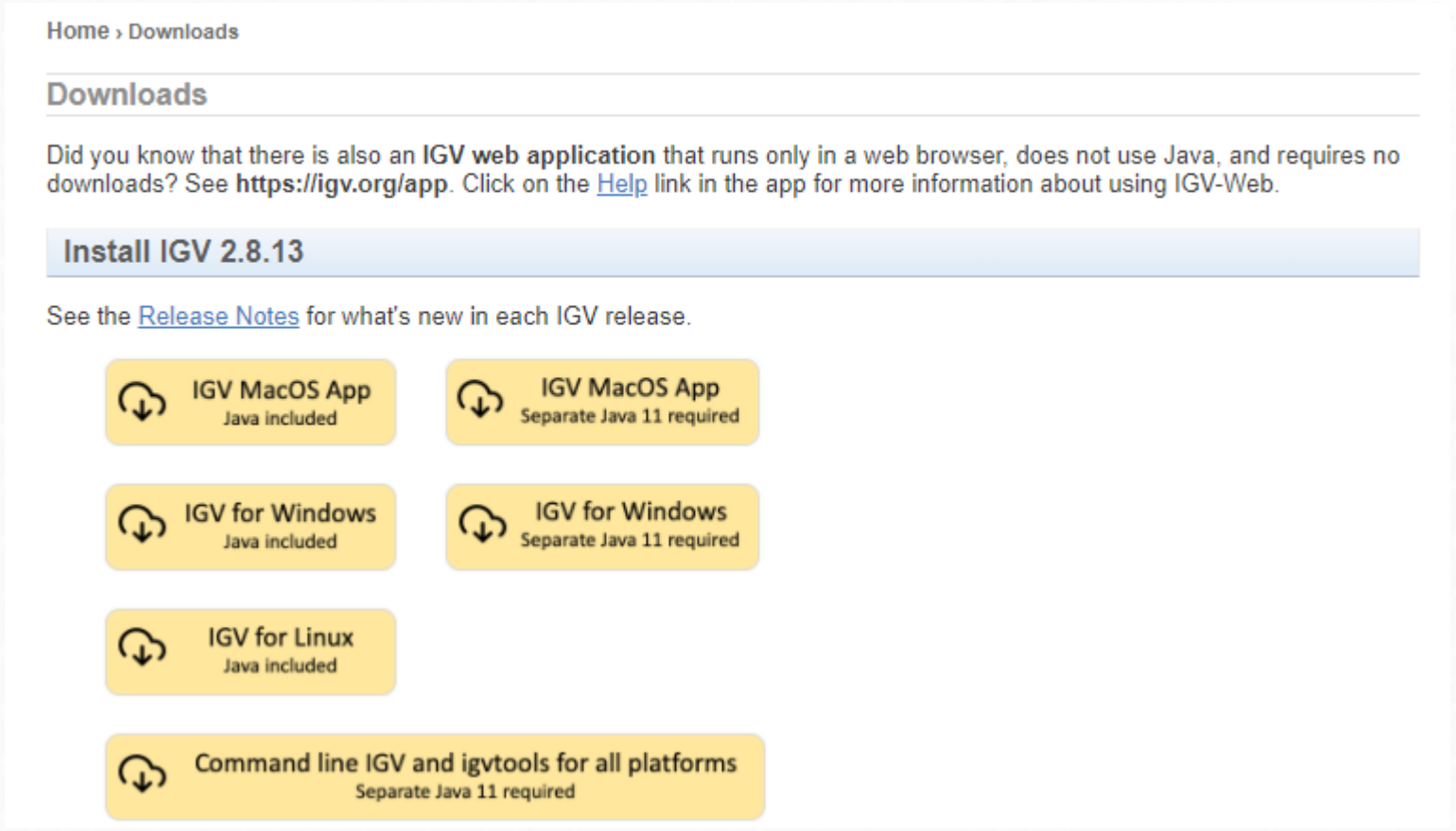
1 Reference sequence identifier
2 Reference sequence length
3 Number of mapped reads
4 Number of placed but unmapped reads
(typically unmapped partners of mapped reads)

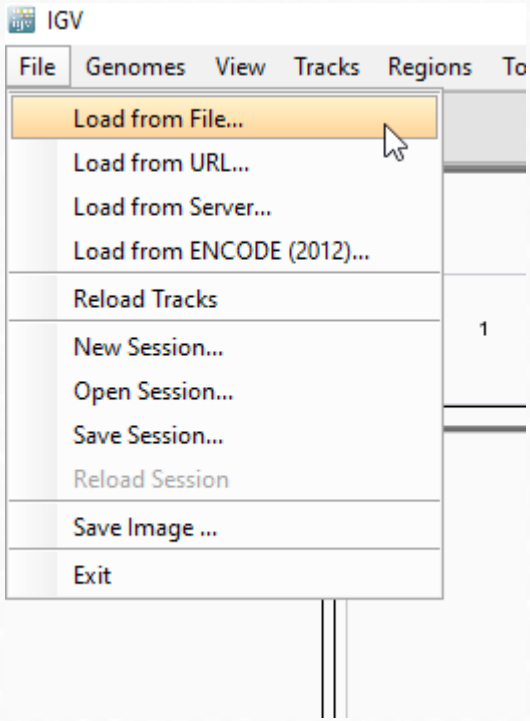
6. Visualise the BAM file with IGV





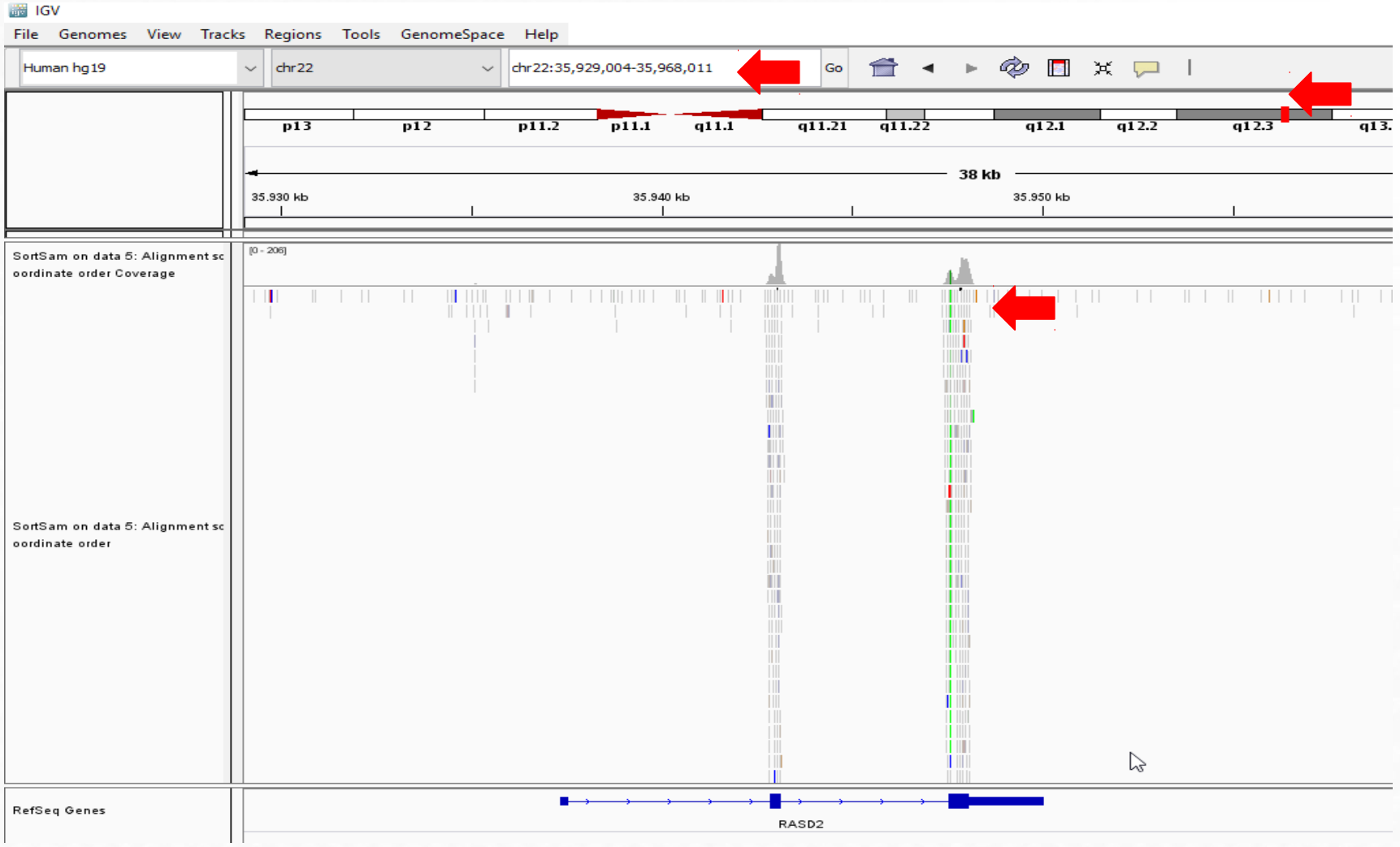
https://software.broadinstitute.org/software/igv/download



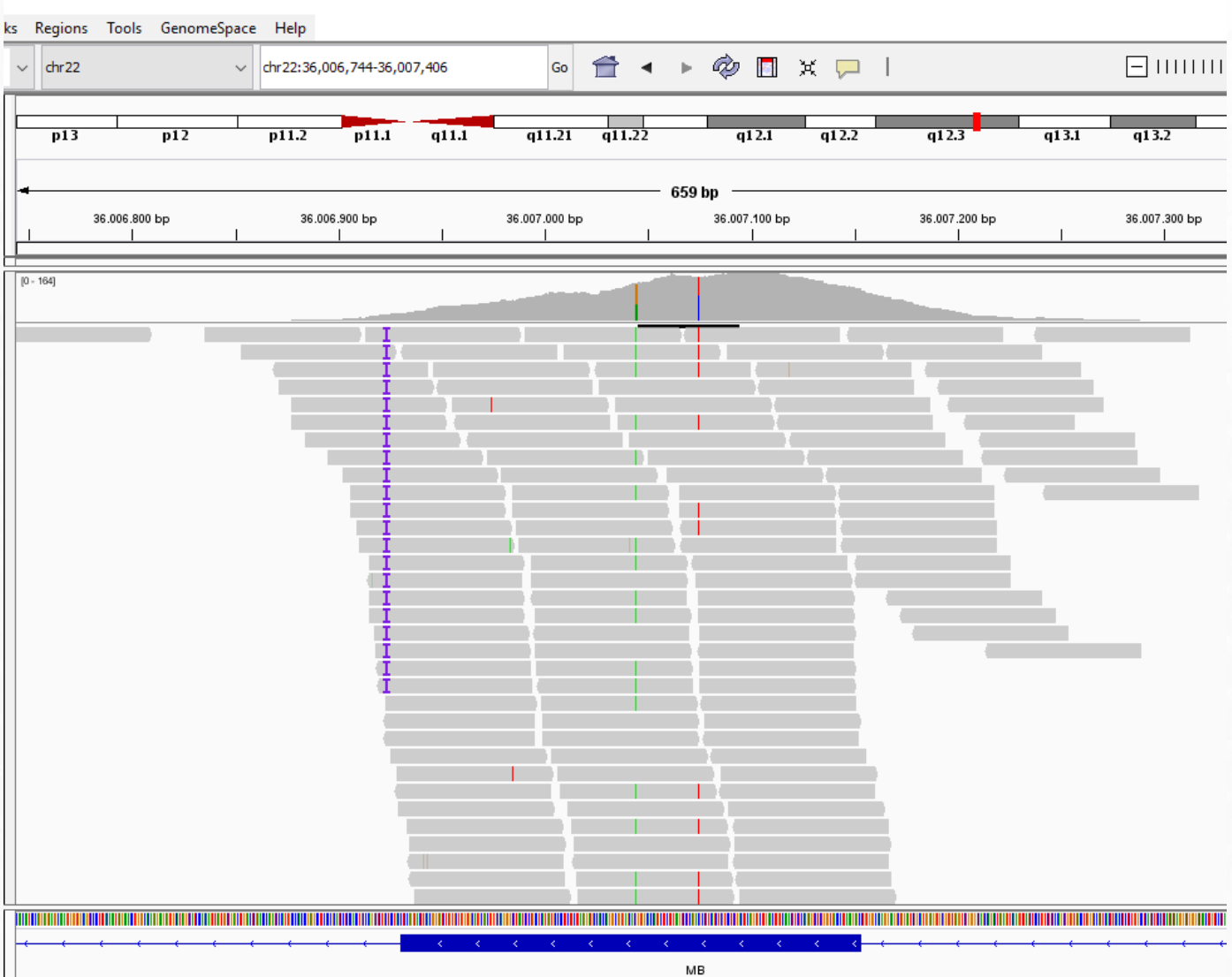


Load the .bam file

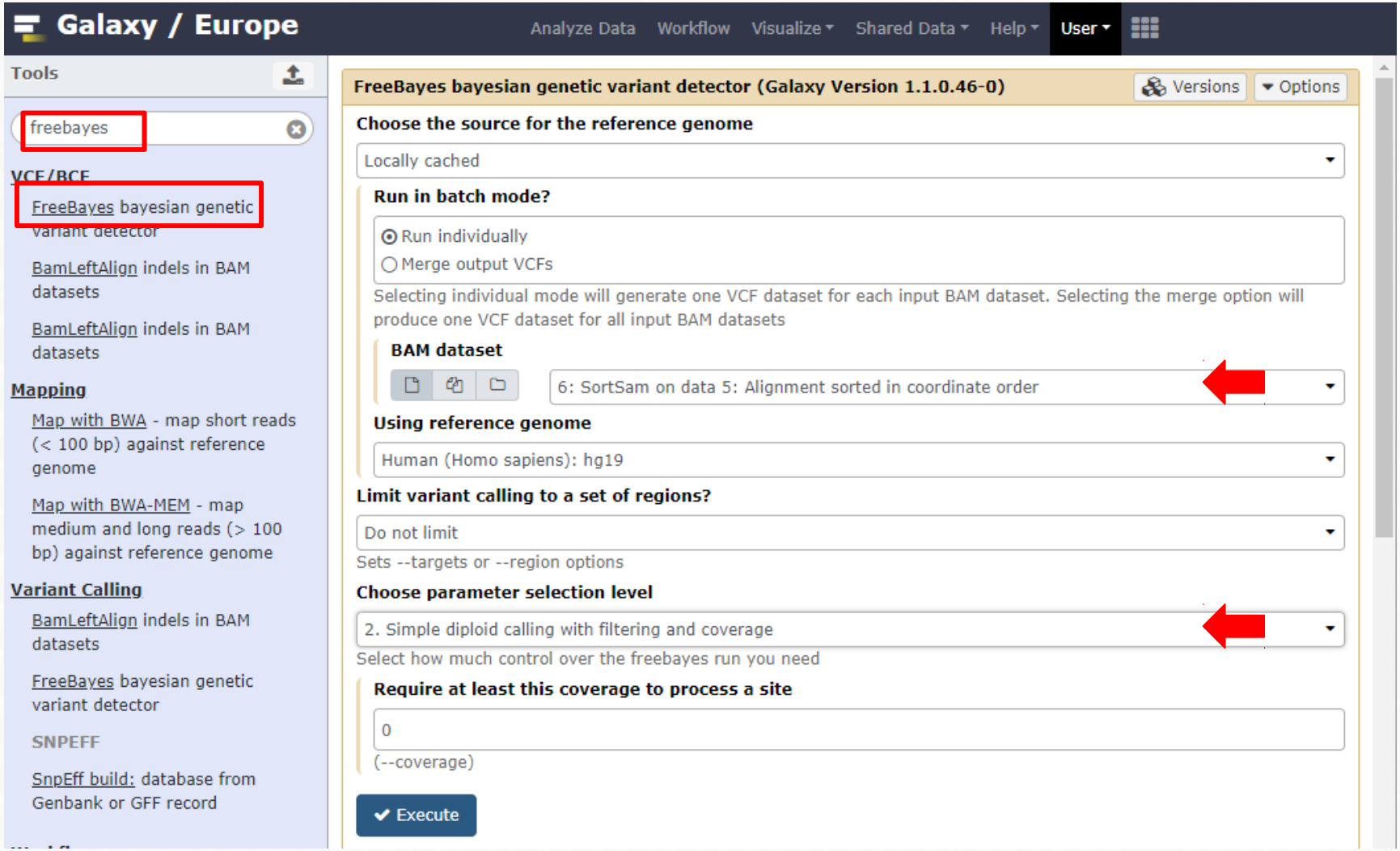
6. Visualize the BAM file with IGV



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



8. Call variants with FreeBayes

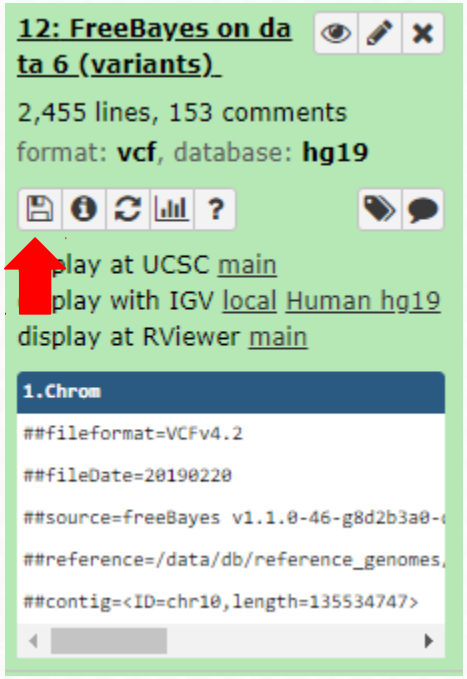


Vall d'Hebron
Institut de Recerca

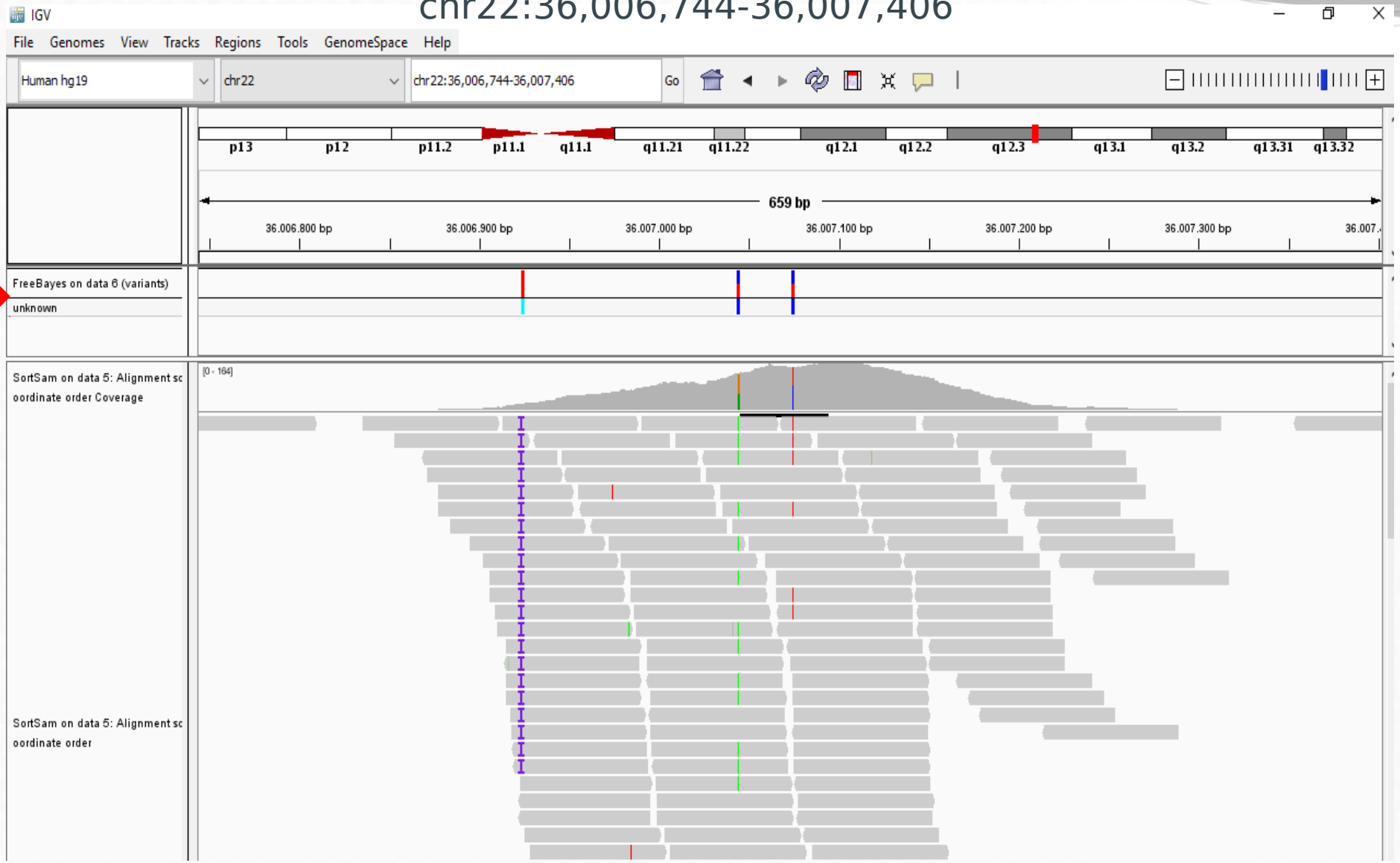
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INF
chr11	116691511	.	GGACAGACAGACAGACAG	GGACAGACAGACAG	3109.94	.	AB
chr11	116691634	.	C	A	267384	.	AB
chr11	116692334	.	C	T	744501	.	AB
chr11	116692694	.	G	A	6855.03	.	AB
chr11	116693464	.	C	T	5143.08	.	AB
chr11	116697848	.	G	A	41.8186	.	AB
chr11	116701535	.	T	C	1782.38	.	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	.	GTTTTTTTTTTG	GTTTTTTTTTTTG	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.

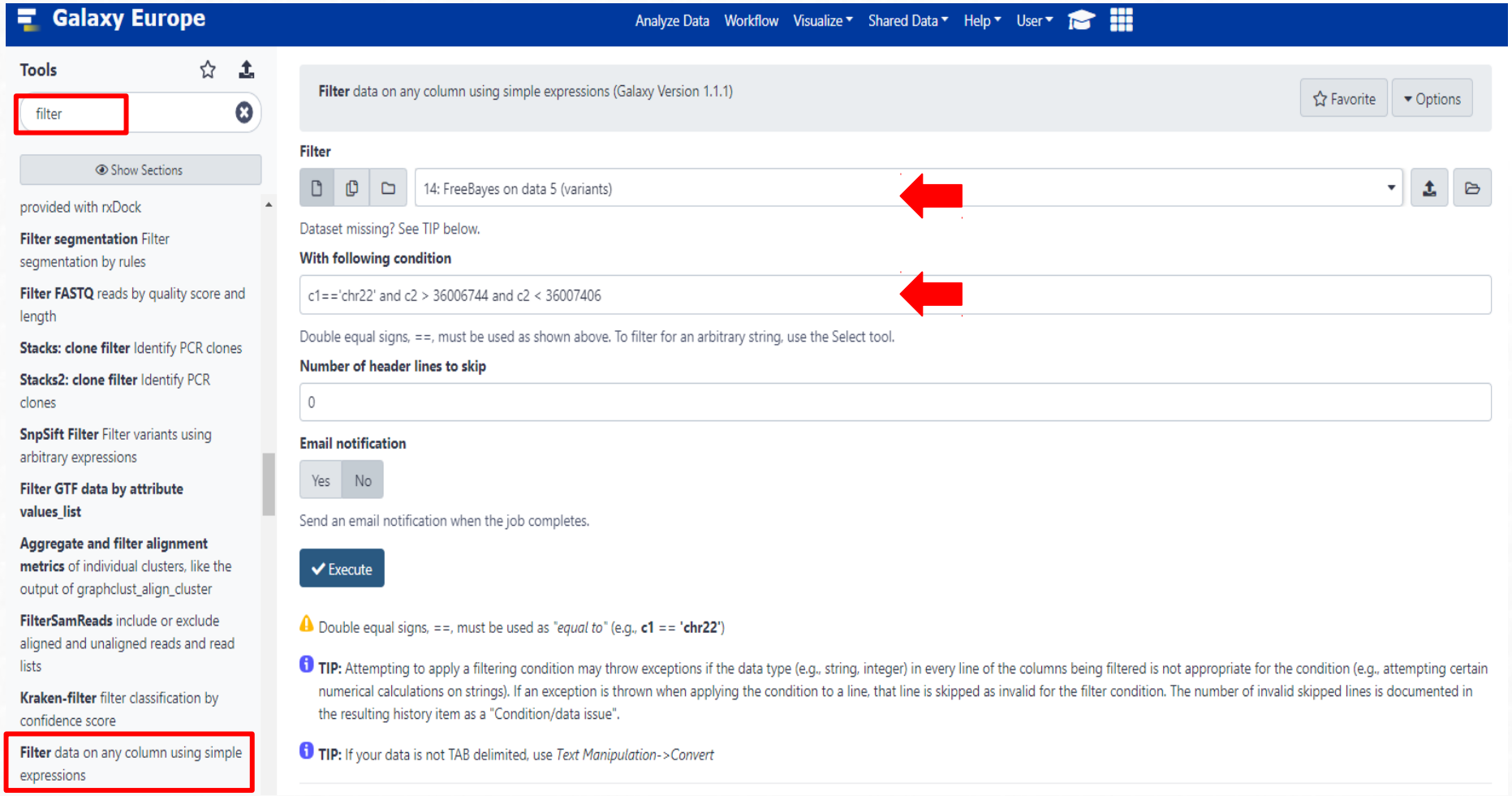
- Visualise the VCF file with IGV



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

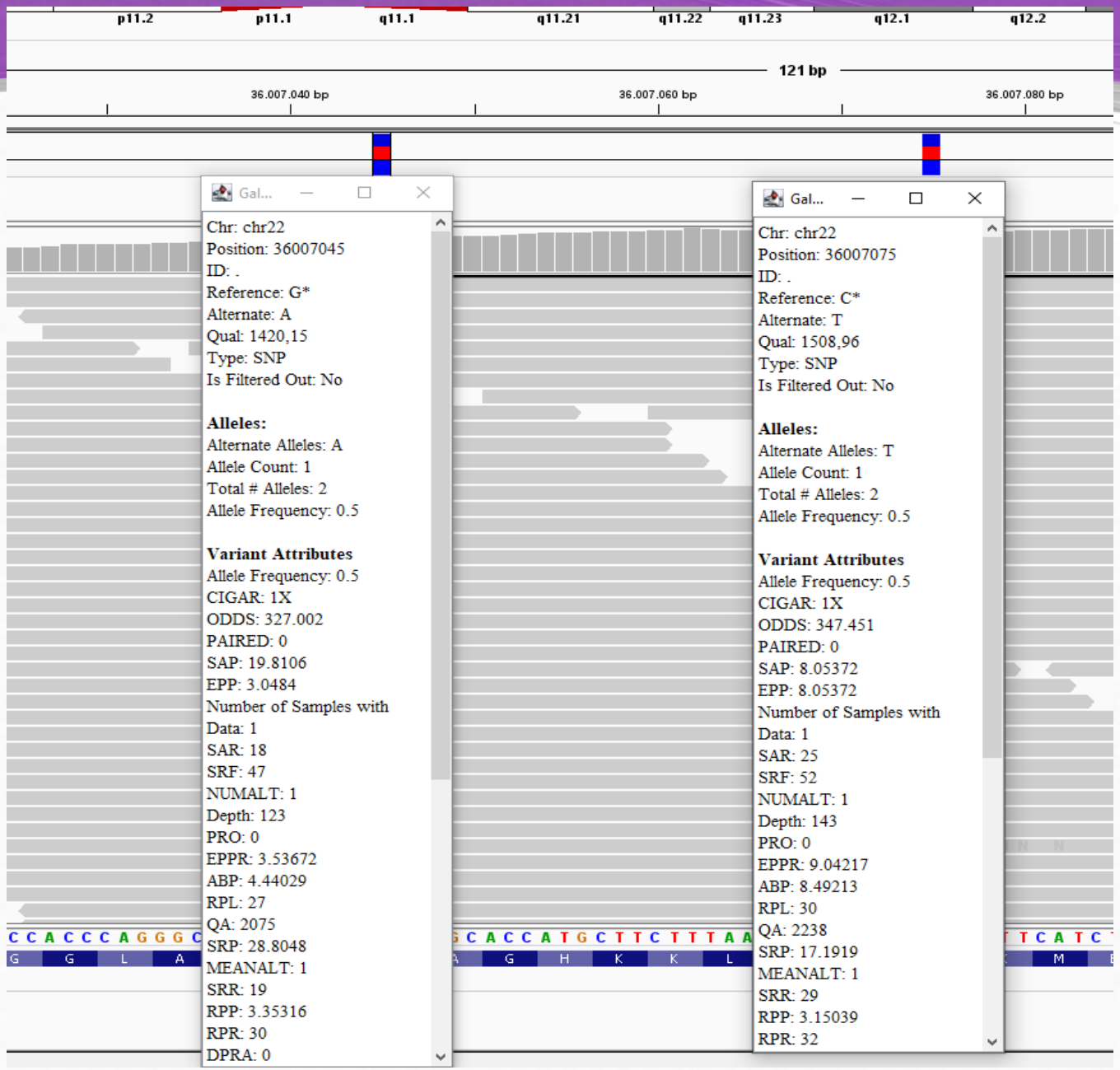


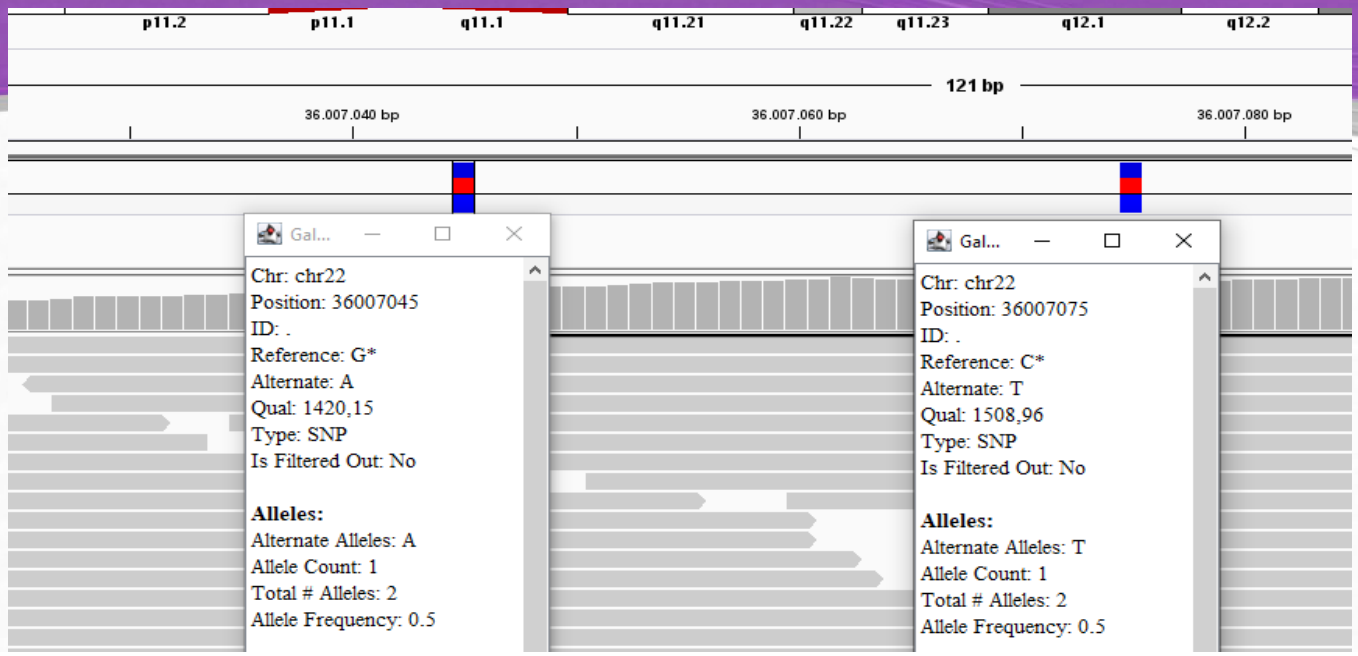
- Filter VCF file



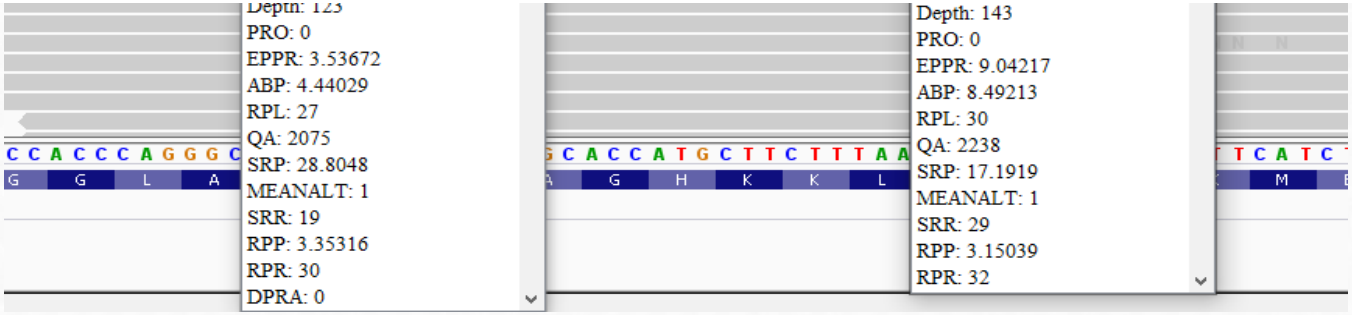
c1=='chr22' and c2 > 36006744 and c2 < 36007406

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=1M1I2M;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;M
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;M


p11.2
p11.1
q11.1
q11.21
q11.22
q11.23
q12.1
q12.2
121 bp
36.007.040 bp
36.007.060 bp
36.007.080 bp
Gal...
Chr: chr22
Position: 36007045
ID: .
Reference: G*
Alternate: A
Qual: 1420,15
Type: SNP
Is Filtered Out: No
Alleles:
Alternate Alleles: A
Allele Count: 1
Total # Alleles: 2
Allele Frequency: 0.5
Variant Attributes
Allele Frequency: 0.5
CIGAR: 1X
ODDS: 327.002
PAIRED: 0
SAP: 19.8106
EPP: 3.0484
Number of Samples with Data: 1
SAR: 18
SRF: 47
NUMALT: 1
Depth: 123
PRO: 0
EPPR: 3.53672
ABP: 4.44029
RPL: 27
QA: 2075
SRP: 28.8048
MEANALT: 1
SRR: 19
RPP: 3.35316
RPR: 30
DPRA: 0
Gal...
Chr: chr22
Position: 36007075
ID: .
Reference: C*
Alternate: T
Qual: 1508,96
Type: SNP
Is Filtered Out: No
Alleles:
Alternate Alleles: T
Allele Count: 1
Total # Alleles: 2
Allele Frequency: 0.5
Variant Attributes
Allele Frequency: 0.5
CIGAR: 1X
ODDS: 347.451
PAIRED: 0
SAP: 8.05372
EPP: 8.05372
Number of Samples with Data: 1
SAR: 25
SRF: 52
NUMALT: 1
Depth: 143
PRO: 0
EPPR: 9.04217
ABP: 8.49213
RPL: 30
QA: 2238
SRP: 17.1919
MEANALT: 1
SRR: 29
RPP: 3.15039
RPR: 32



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;M(



9. Annotating variants with SnpEff

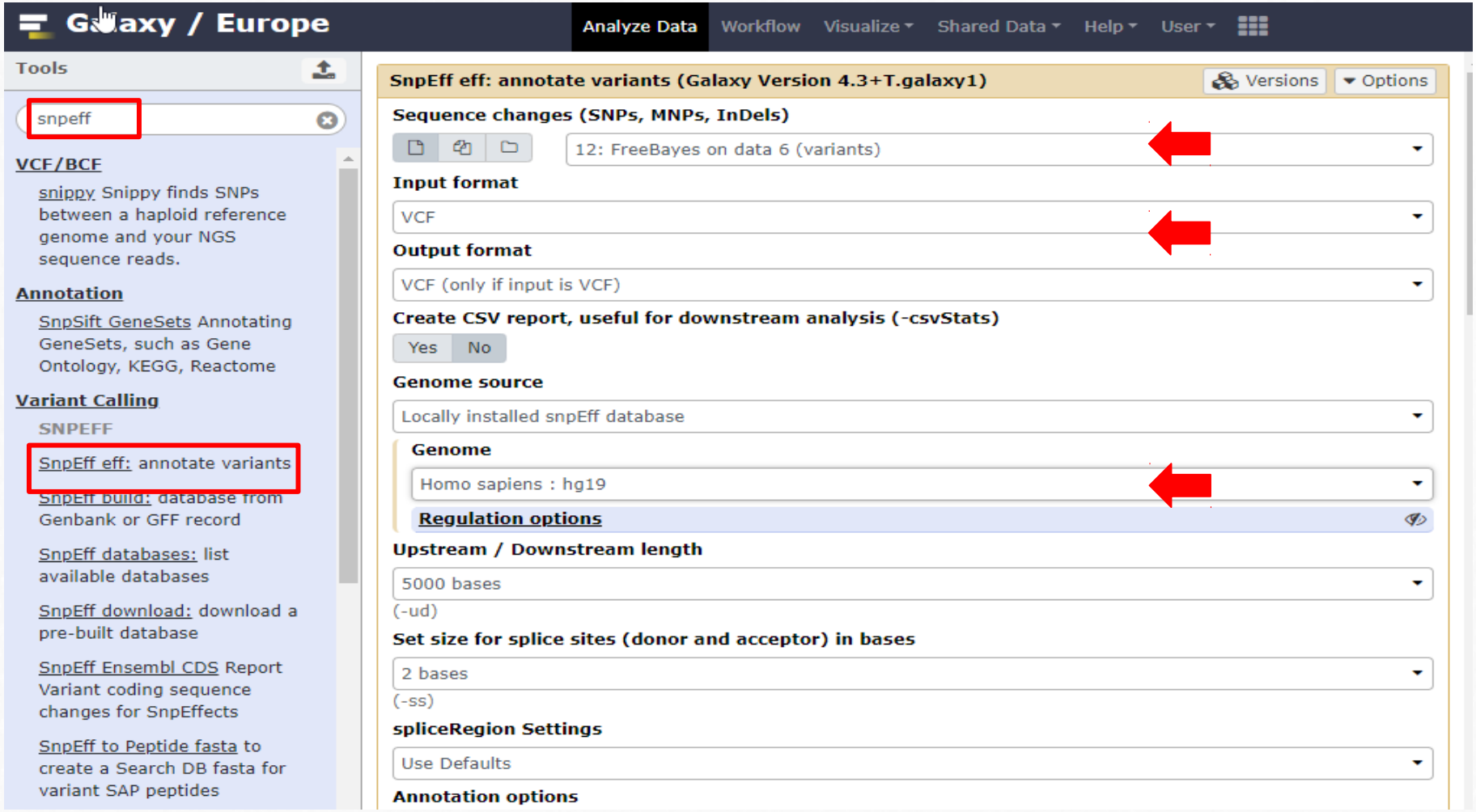
Genetic variant annotation and functional effect prediction toolbox. It annotates and predicts the effects of genetic variants on genes and proteins (such as amino acid changes).

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



1. Hands-On Exome Variant Analysis

Vall d'Hebron
Institut de Recerca

SnpEff will generate two outputs:

- an annotated VCF file
- an HTML report

SnpEff: Variant analysis

Contents

Summary

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/galax
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%