

HANDS ON: INTRODUCTION TO VARIANT ANALYSIS

Bioinformàtica per a la Recerca Biomèdica

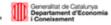
Mireia Ferrer¹, Álex Sánchez^{1,2} Esther Camacho¹, Angel Blanco^{1,2}

1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR 2 Departament de Genètica, Microbiologia i Estadística, UB





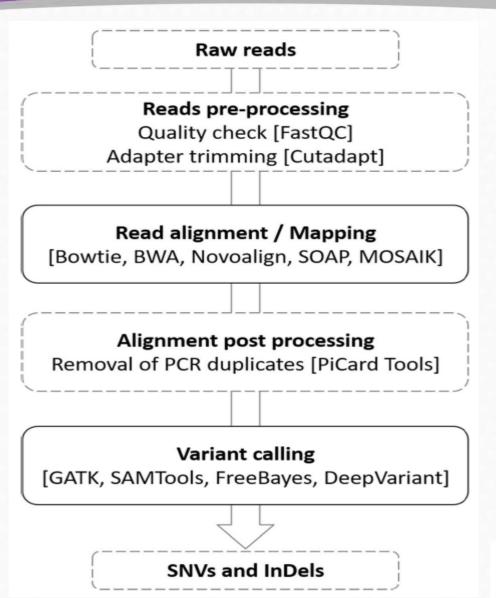












Kumaran *et al. BMC Bioinformatics* (2019) 20:342 https://doi.org/10.1186/s12859-019-2928-9



This presentation is mainly based in a tutorial published by:



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 Virtua and these are used in our formal workshops and are also available for use to delive workshops or for self-directed learning.



genomics

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/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_calling_galaxy_1/variant_



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:

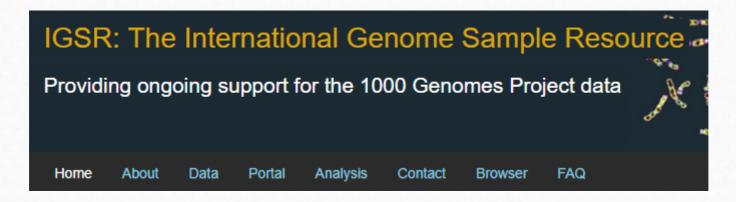
Detecting small variants in human genomic DNA using a small set of reads from chromosome 22



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







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

https://usegalaxy.eu/join-training/ueb_bi2021

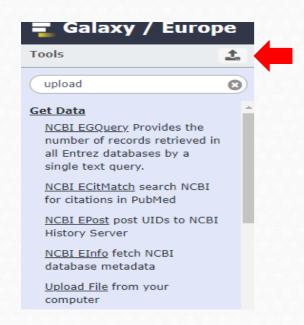


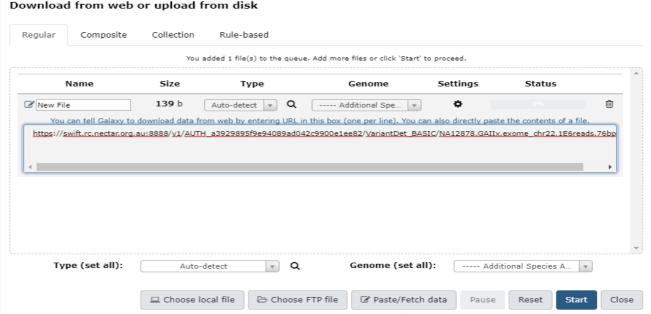


Import data for the tutorial

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



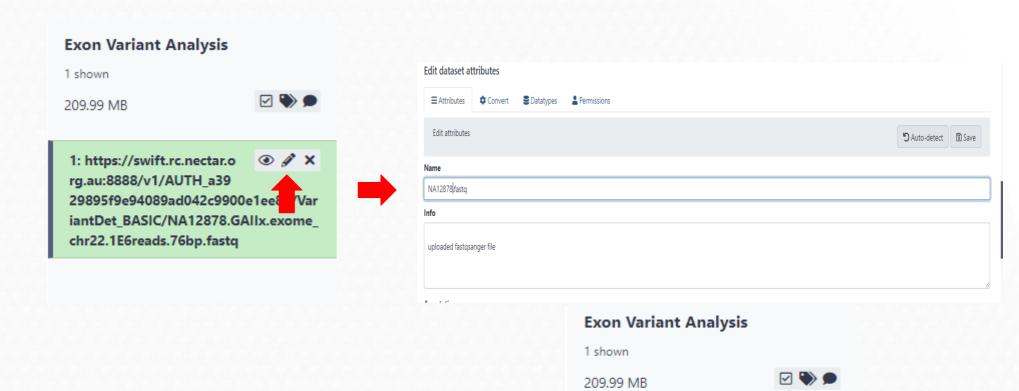






1: NA12878.fastq

Change the name of file:

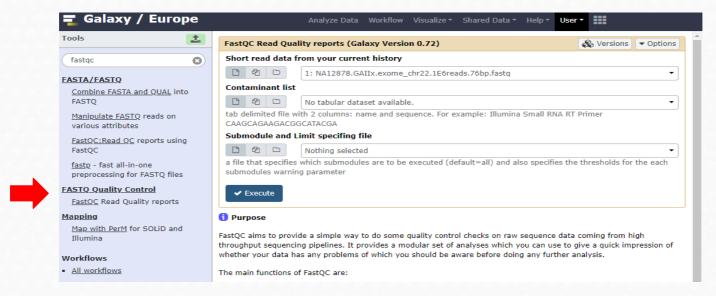




Take a look at the FASTQ file

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

3. Assessing read quality from the FASTQ files





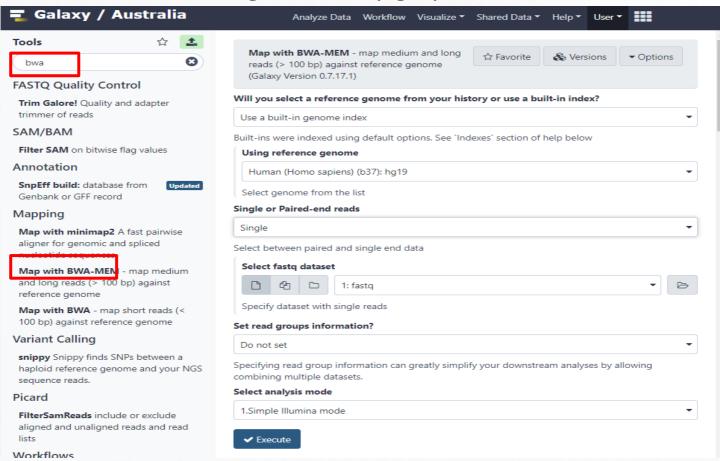
Results of the Quality Control





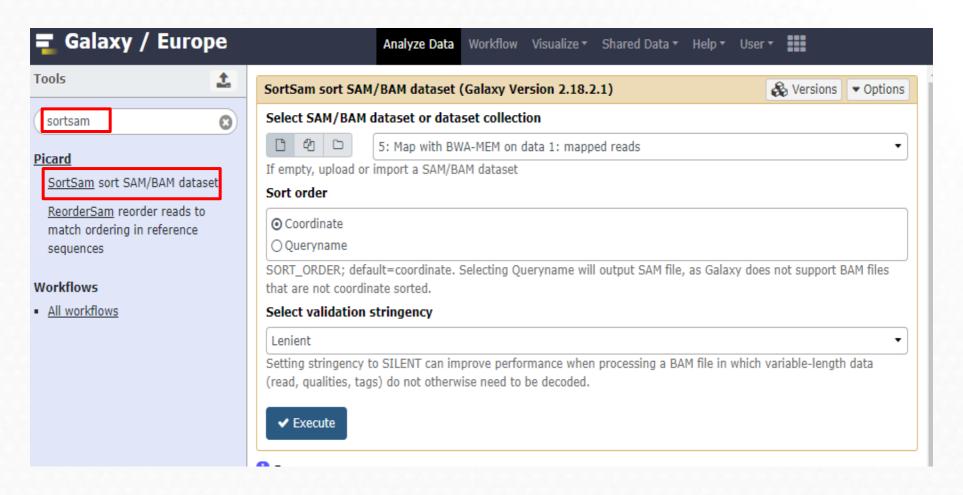
4. Align the reads with BWA

Reference Genome: *Human genome 19 (hg19)*





Sort the BAM file



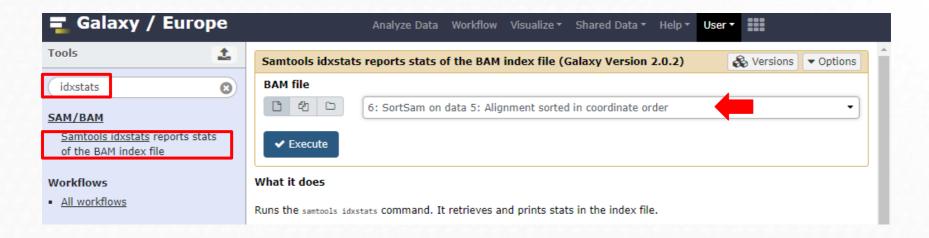


61CC3AAXX100125:/:118:2538:55//	16	chr22	16050954
61CC3AAXX100125:7:1:17320:13701	0	chr22	16052274
61CC3AAXX100125:7:93:5100:14497	0	chr14	19790076
61CC3AAXX100125:6:92:7549:15004	16	chr22	16052936
61CC3AAXX100125:5:7:1488:7780	16	chr22	16053177
61CC3AAXX100125:7:72:14903:20386	16	chr22	16053702
61CC3AAXX100125:7:88:9942:19183	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



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
chr22	51304566	1101584	0
chr2	243199373	4094	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
			_

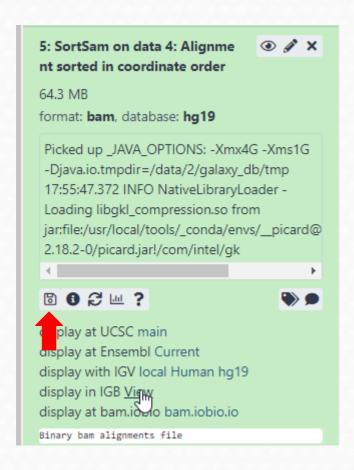
Column Description

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

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.

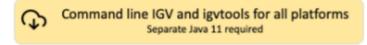




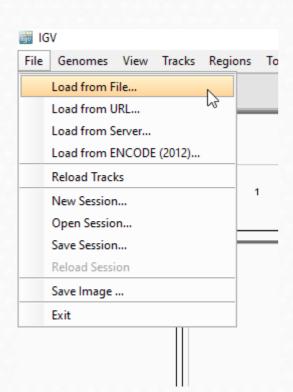








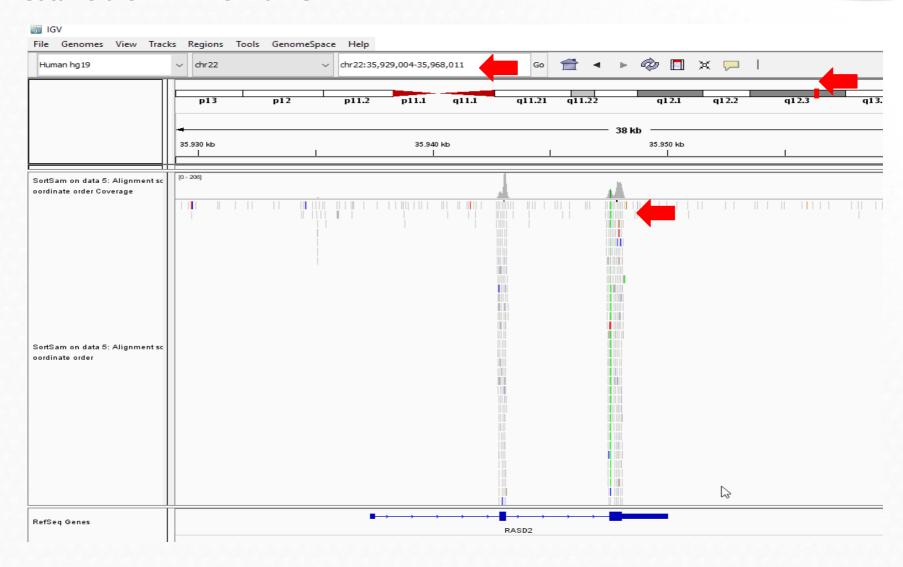




Load the .bam file

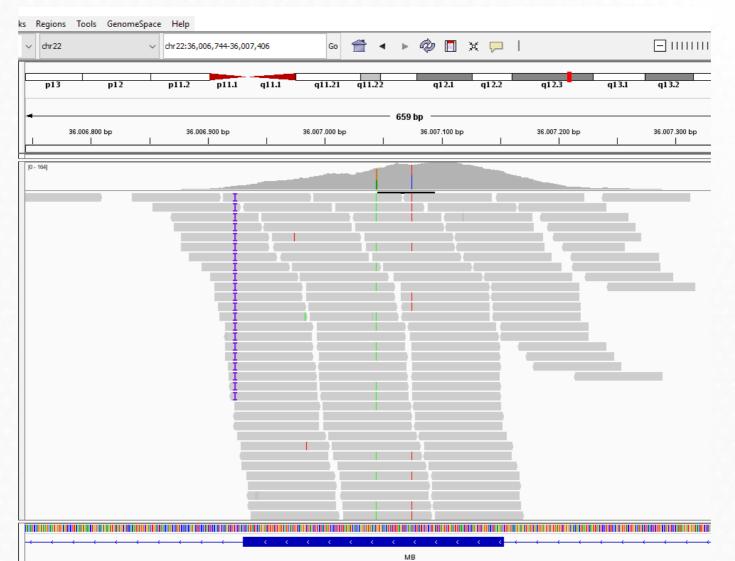


6. Visualize the BAM file with IGV



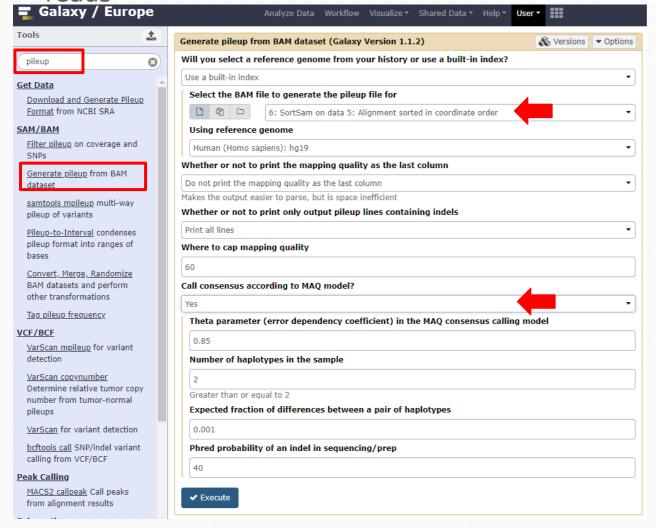


chr22:36,006,744-36,007,406





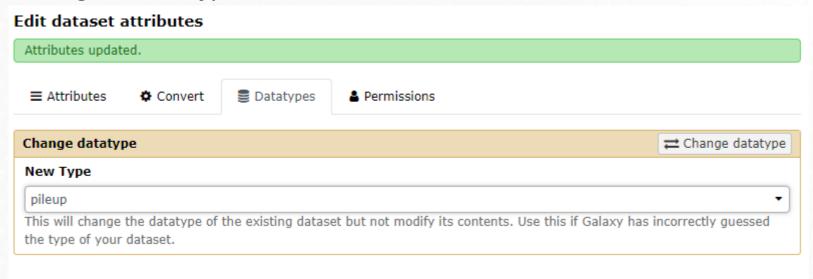
7. Generate a pileup file: a column-wise representation of the aligned reads



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.



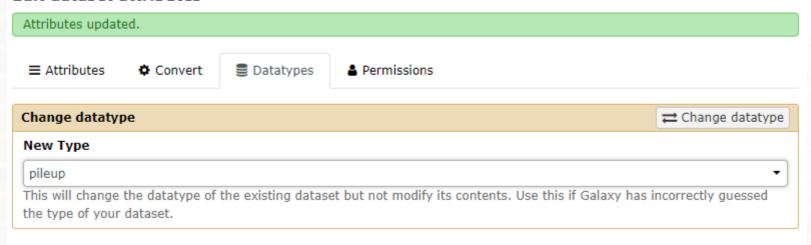
Change file datatype

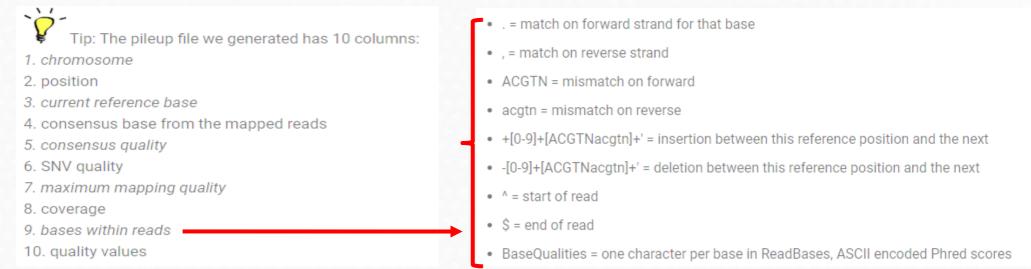




Change file datatype

Edit dataset attributes







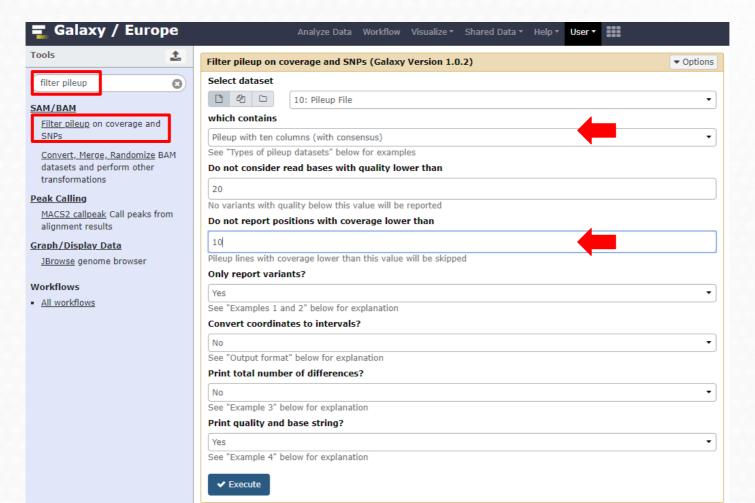
Filter the pileup file

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

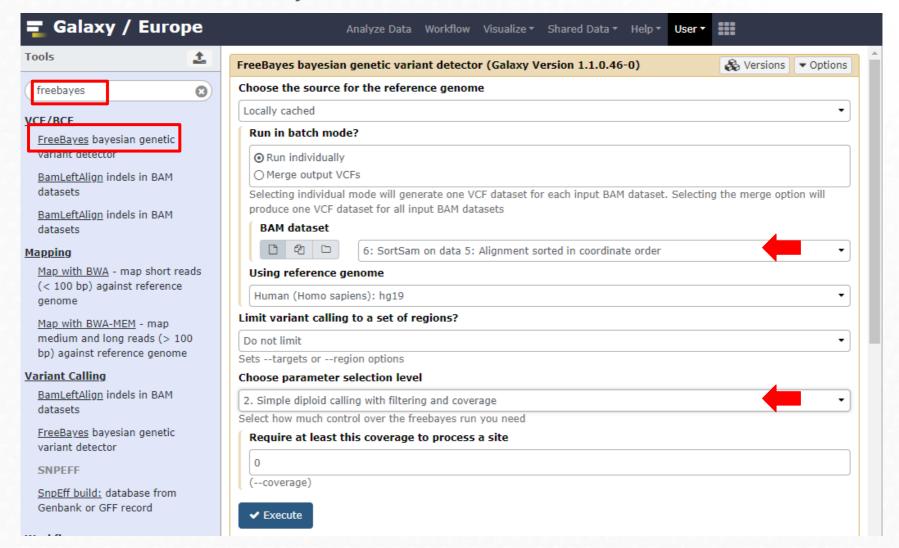




chr21	15441242	Α	Α	187	0	19	62	Т,.,,
chr21	15441261	Т	Т	32	0	5	16	CC,,,C,cCcc,.CCC
chr21	15441264	Т	Т	11	0	5	16	CCcccCcccCccCCCC
chr21	48110986	С	С	53	0	0	16	,.,,.,,,,.g
chr22	16084443	G	G	56	0	0	18	,,,,,,,,,,,,,,a
chr22	16084718	G	G	27	0	0	10	A.A.AAAA
chr22	16084749	G	G	41	0	0	23	T,,,^!.^!.^!.
chr22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



8. Call variants with FreeBayes



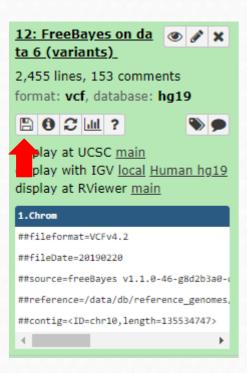
#CHROM	POS I	ID I	REF	ALT	OUAL	FILTER INF
chr11	116691511.		GGACAGACAGACAG	GGACAGACAGACAG	3109.94	. AB
chr11	116691634.		C	A	267384	
chr11	116692334.	. (Č	T	744501	
chr11	116692694.	.	Ğ	Â	6855.03	. AB
chr11	116693464.	. (Č	T	5143.08	. AB
chr11	116697848.		Ğ	A	41.8186	. AB
chr11	116701535.	.	Ť	C	1782.38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



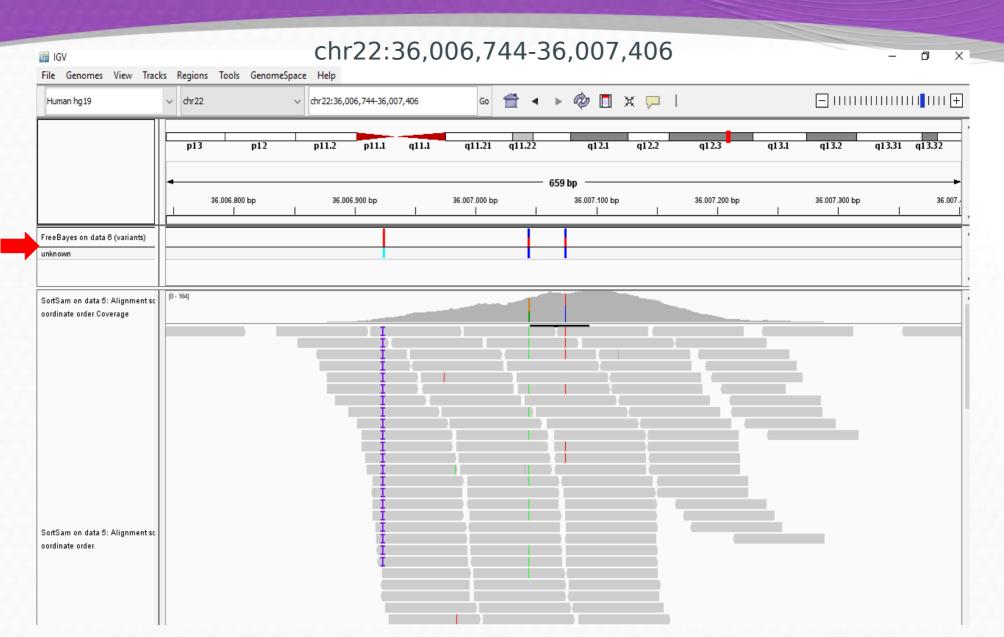
Col	Field	Description
1	CHROM	Chromosome name
2	POS	1-based position. For an indel, this is the position preceding the indel.
3	ID	Variant identifier (optional). Usually the dbSNP rsID.
4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
5	ALT	Comma delimited list 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

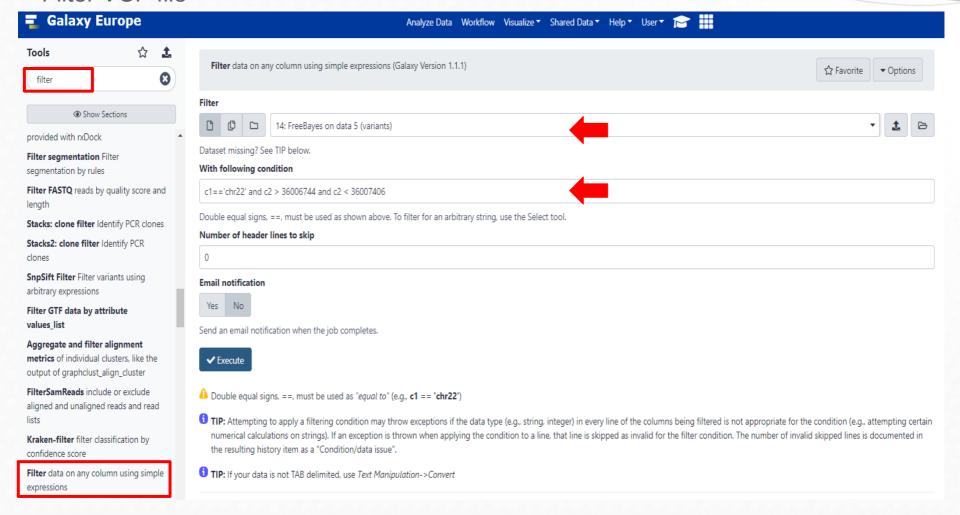








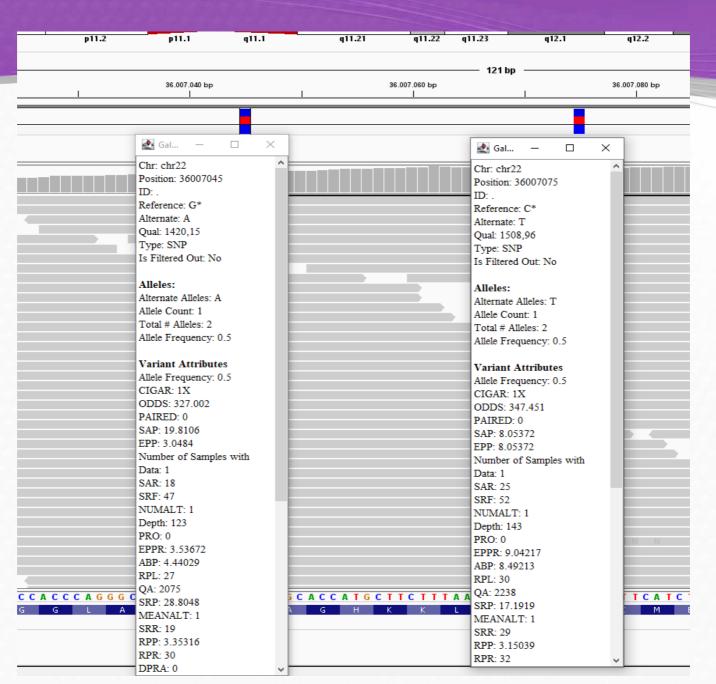
Filter VCF file



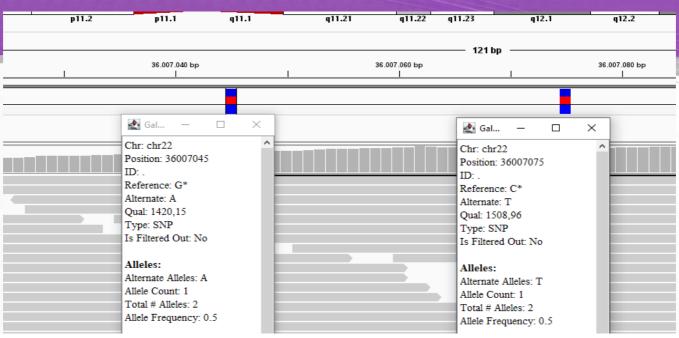


Chrom	Pos	ID	Ref	Alt	Qual	Filter	Info
chr22	36006923		GCT	GCCT	645.607		AB = 0; ABP = 0; AC = 2; AF = 1; AN = 2; AO = 21; CIGAR = 1M112M; DP = 21; DPB = 28; DPRA = 0; EPP = 5.59539; EPPR = 0; GTI = 0; LEN = 1; MEANALT = 1; MQM = 60; MQMR = 0; DPRA = 0; DPR
chr22	36007045		G	Α	1420.15		AB=0.463415;ABP=4.44029;AC=1;AF=0.5;AN=2;AO=57;CIGAR=1X;DP=123;DPB=123;DPRA=0;EPP=3.0484;EPPR=3.53672;GTI=0;LEN=1;MEANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MCANALT=1;MC
chr22	36007075		С	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; NEANALT = 1; NE

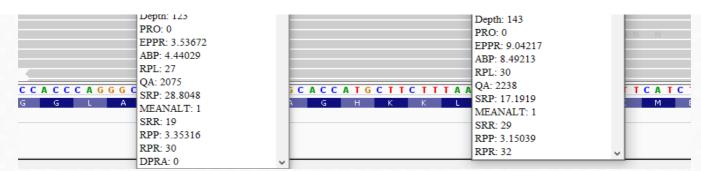








Chrom	Pos	ID	Ref	Alt	Qual	Filter	Info
chr22	36006923		GCT	GCCT	645.607		AB=0;ABP=0;AC=2;AF=1;AN=2;AO=21;CIGAR=1M1I2M;DP=21;DPB=28;DPRA=0;EPP=5.59539;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;N
chr22	36007045	ı	G	А	1420.15		AB=0.463415;ABP=4.44029;AC=1;AF=0.5;AN=2;AO=57;CIGAR=1X;DP=123;DPB=123;DPRA=0;EPP=3.0484;EPPR=3.53672;GTI=0;LEN=1;MEANALT=1;MQ
chr22	36007075		C	Ţ	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;MC





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).

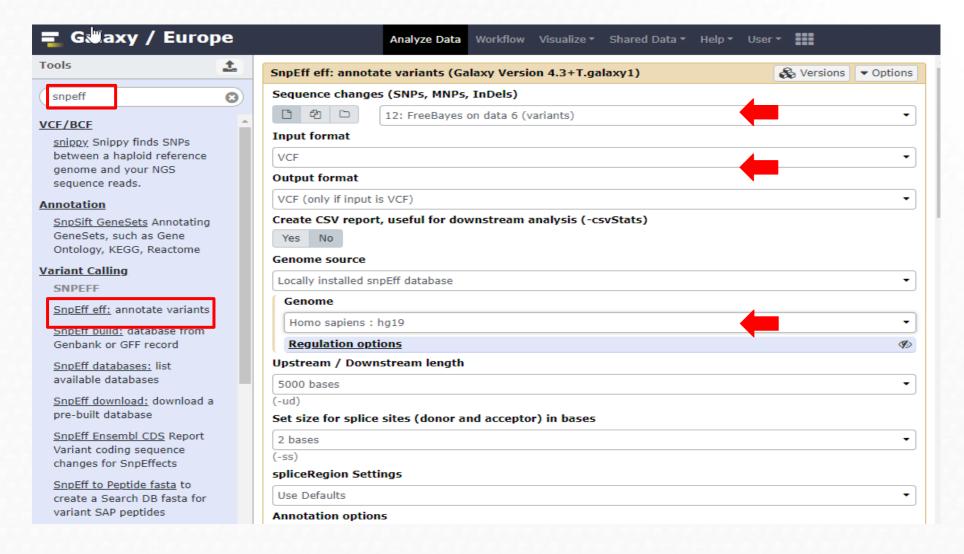
http://snpeff.sourceforge.net/

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



9. Annotating variants with SnpEff



1. Hands-On Exome Variant Analysis



SnpEff will generate two outputs:

an annotated VCF file

an HTML report

SnpEff: Variant analysis

Contents

Summary Variant rate by chromosome Variants by type Number of variants by impact Number of variants by functional class Number of variants by effect Quality histogram InDel length histogram Base variant table Transition vs transversions (ts/tv) Allele frequency Codon change table <u>Amino acid change table</u>

7.255

3,137,161,265

1 variant every 253,813 bases

657,375,723

Chromosome variants plots

Number of effects Genome total length

Variant rate

Genome effective length

.*3740T>C|||||3055|,C|intron_variant|MODIFIER|POLRMT|POLRMT|transcript|NM_005035.3|protein_coding|11/20|c.2764c.*3754A>C|||||3069|,C|intron_variant|MODIFIER|POLRMT|POLRMT|transcript|NM_005035.3|protein_coding|11/20|c.2763

T|POLRMT|transcript|NM 005035.3|protein coding|13/20|c.3066+12A>C|||||

|||||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|||||

)|,A|intron_variant|MODIFIER|POLRMT|POLRMT|transcript|NM_005035.3|protein_coding|14/20|c.3154-29C>T|||||

5||||2249|,G|intron_variant|MODIFIER|POLRMT|POLRMT|transcript|NM_005035.3|protein_coding|13/20|c.3067-112T>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

F>A|p.A|a933A|a|2855/3800|2799/3693|933/1230||,T|downstream_gene_variant|MODIFIER|HCN2|HCN2|transcript|NM_00

27A>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|||||

n_coding|11/21|c.2747A>C|||||,G|structural_interaction_variant|HIGH|POLRMT|POLRMT|interaction|4BOC:A_827-A_916:N '0/3800|2714/3693|905/1230||,C|downstream_gene_variant|MODIFIER|HCN2|HCN2|transcript|NM_001194.3|protein_codii

/3800|2699/3693|900/1230||,C|downstream_gene_variant|MODIFIER|HCN2|HCN2|transcript|NM_001194.3|protein_coding 35.3|protein_coding|11/21|c.2674T>G||||||,C|structural_interaction_variant|HIGH|POLRMT|POLRMT|interaction|3SPA:A_81

Summary

-59A>G|p.Glu890Gly|2725/3800|2669/3693|890/1230||,C|downstream_gene_variant|MODIFIER|HCN2|HCN2|transcript|NM_ Genome 2020-12-01 21:42 Date SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani SnpEff version Command line arguments o vcf -stats /data/dnb03/galaxy db/job working directory/013/520/13520986/outputs/galaxy-Warnings 0 2.583 Number of lines (input file) 2,590 Number of variants (before filter) Number of not variants 0 (i.e. reference equals alternative) Number of variants processed 2,590 (i.e. after filter and non-variants) Number of known variants 0 (0%) (i.e. non-empty ID) Number of multi-allelic VCF entries (i.e. more than two alleles)



Genome	hg19									
Date	2020-12-01 21:42									
SnpEff version	SnpEff 4.3t (build 2017-11-24 10	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani								
Command line arguments	SnpEff -i vcf -o vcf -stats /da	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 varia	nts by type							
Number of known variants (i.e. non-empty ID)	0 (0%)	Туре	Total							
Number of multi-allelic VCF entries (i.e. more than two alleles)	7	MNP	<mark>-2,389</mark> 61							
Number of effects	7,255	INS	60							
Genome total length	3,137,161,265	DEL	73							

657,375,723

1 variant every 253,813 bases

Genome effective length

Variant rate

Туре	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/dnl	b03/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 Num	aber variants by type
Number of known variants (i.e. non-empty ID)	0 (0%)	Type Total

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

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%