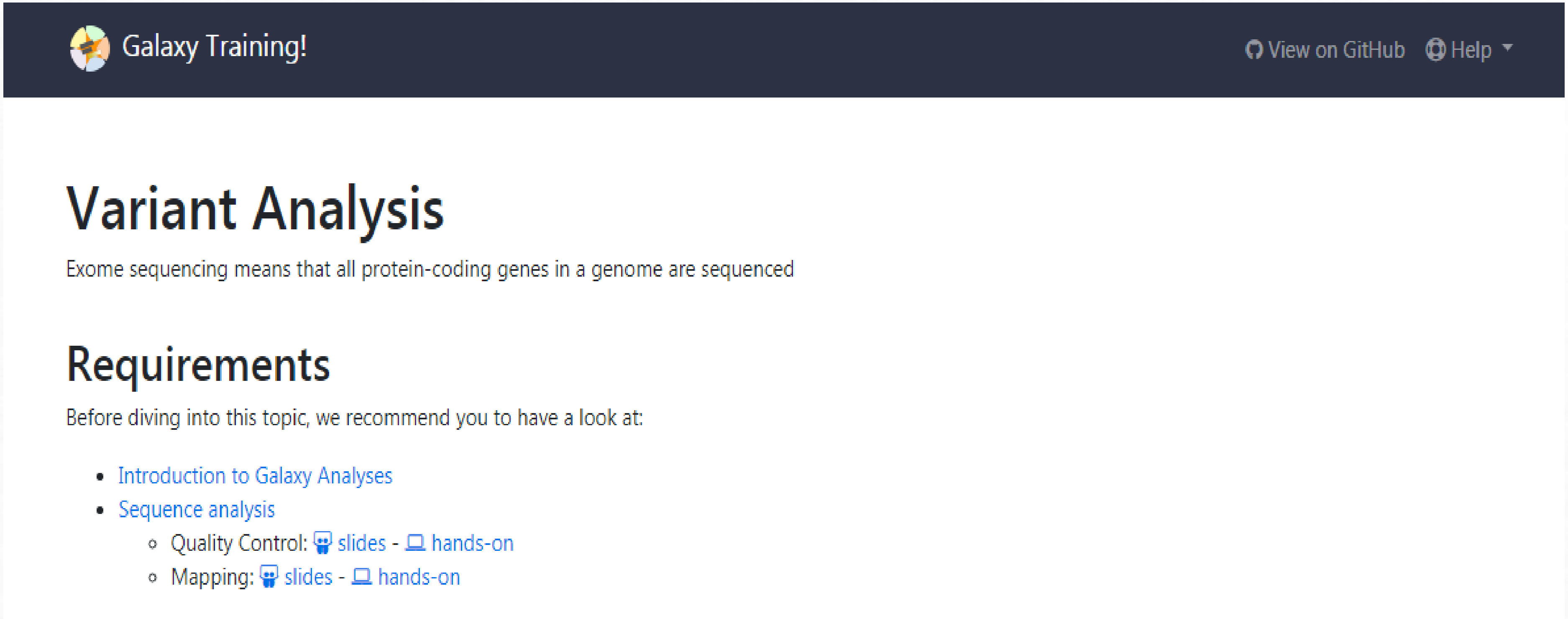
Hands-On. Exome Variant Analysis

Bioinformàtica per a la Recerca Biomèdica
Ricardo Gonzalo Sanz
ricardo.gonzalo@vhir.org
03/12/2018

1. Hands-On Exome Variant Analysis

https://galaxyproject.github.io/training-material/topics/variant-analysis/



Variant Analysis

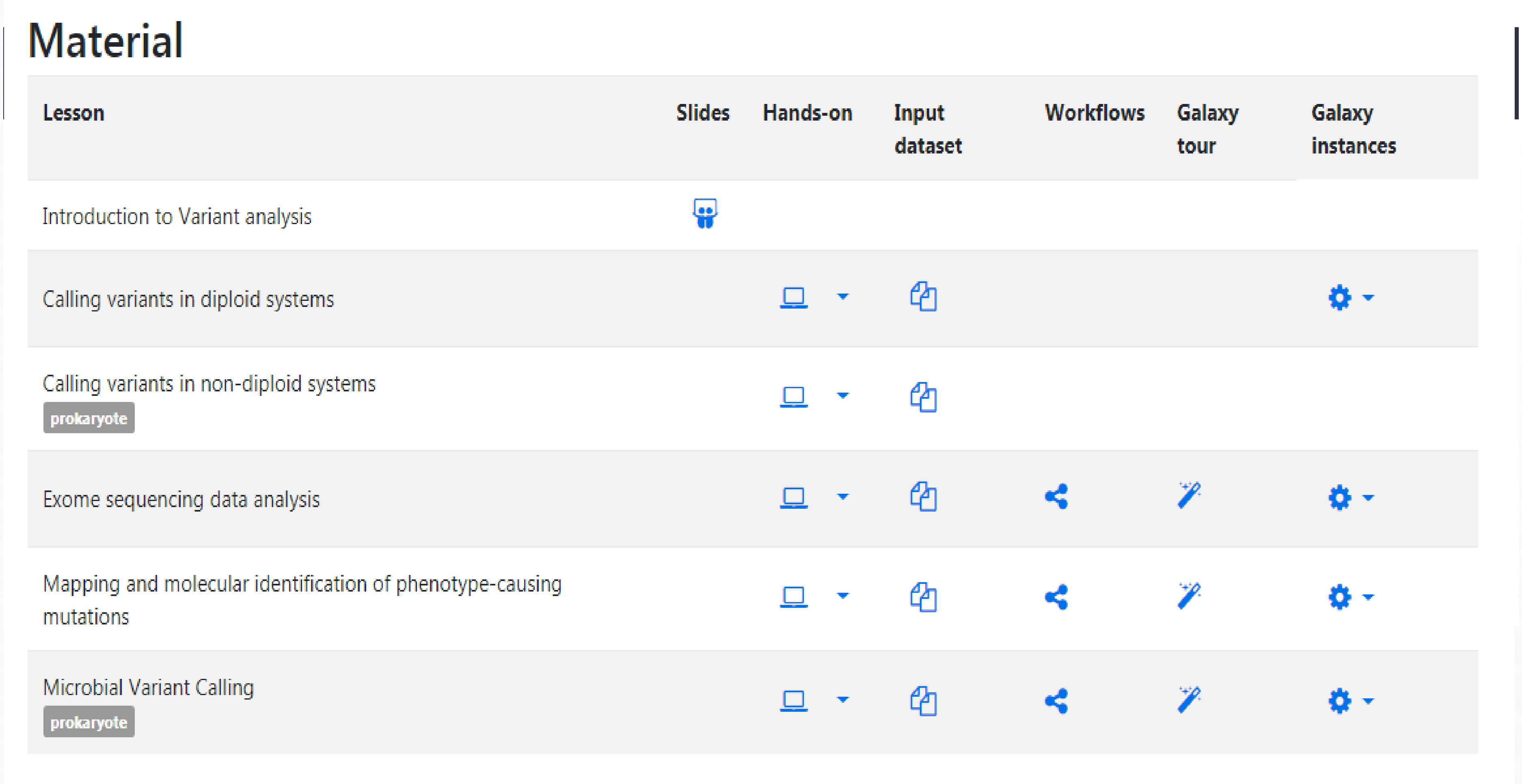
Exome sequencing means that all protein-coding genes in a genome are sequenced

Requirements

Before diving into this topic, we recommend you to have a look at:

- Introduction to Galaxy Analyses
- Sequence analysis
  - Quality Control: slides - hands-on
  - Mapping: slides - hands-on

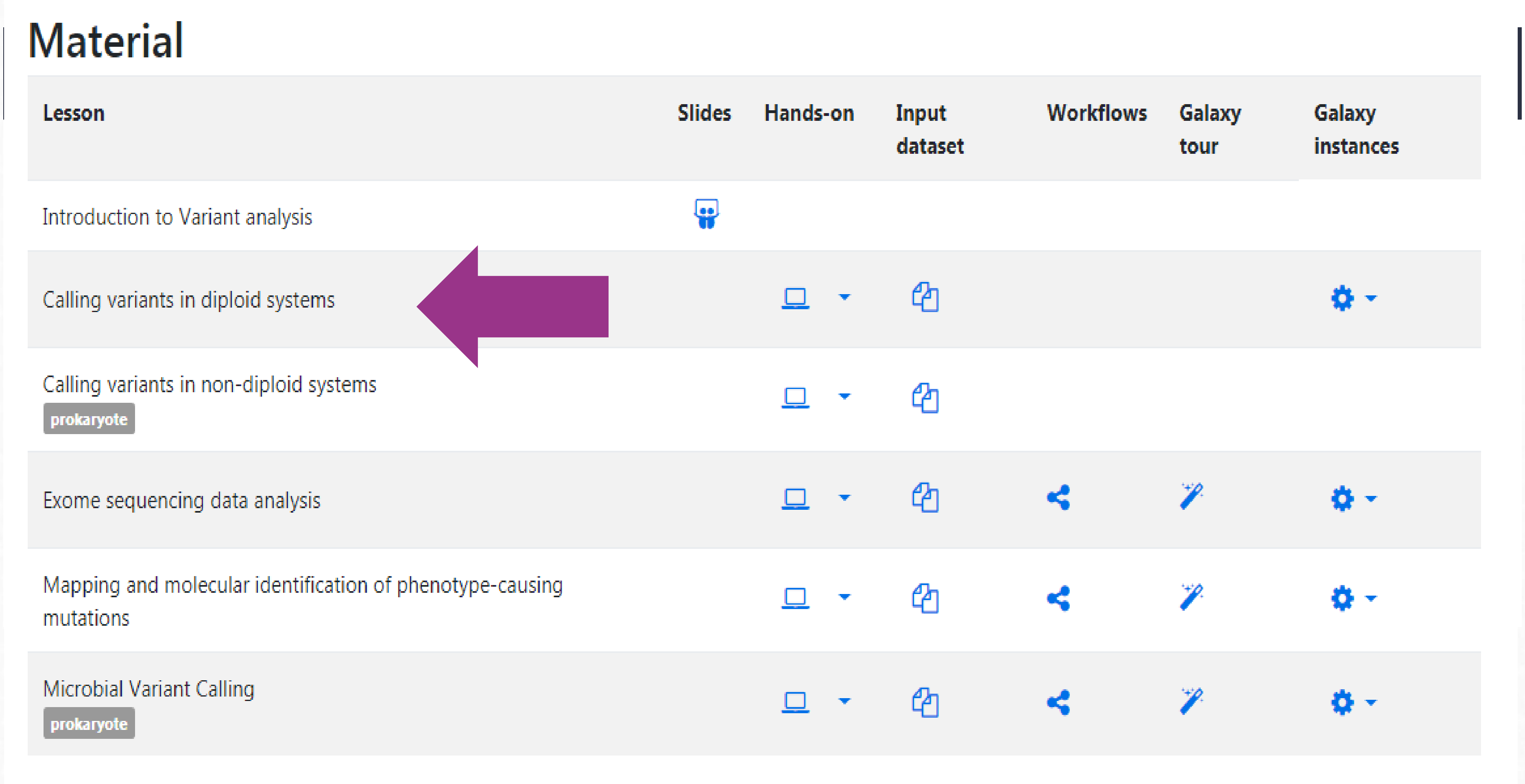
1. Hands-On Exome Variant Analysis



Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to Variant analysis	▣					
Calling variants in diploid systems		💻 ▾	⎘			⚙ ▾
Calling variants in non-diploid systems prokaryote		💻 ▾	⎘			
Exome sequencing data analysis		💻 ▾	⎘	⋖	🪄	⚙ ▾
Mapping and molecular identification of phenotype-causing mutations		💻 ▾	⎘	⋖	🪄	⚙ ▾
Microbial Variant Calling prokaryote		💻 ▾	⎘	⋖	🪄	⚙ ▾

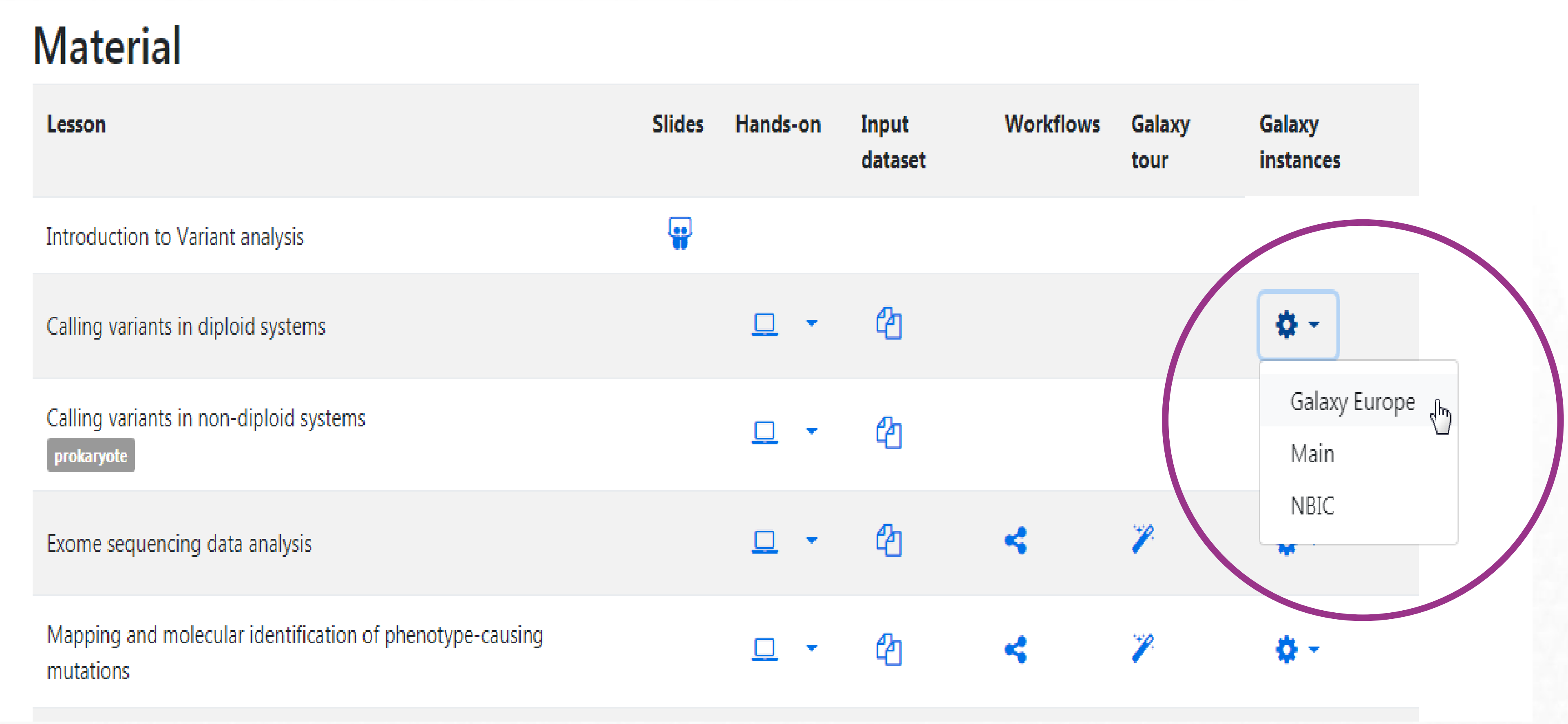
1. Hands-On Exome Variant Analysis



Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to Variant analysis	🖵					
Calling variants in diploid systems		💻 ▾	🗐			⚙ ▾
Calling variants in non-diploid systems prokaryote		💻 ▾	🗐			
Exome sequencing data analysis		💻 ▾	🗐	⋖	🪄	⚙ ▾
Mapping and molecular identification of phenotype-causing mutations		💻 ▾	🗐	⋖	🪄	⚙ ▾
Microbial Variant Calling prokaryote		💻 ▾	🗐	⋖	🪄	⚙ ▾

1. Hands-On Exome Variant Analysis



Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to Variant analysis	📽					
Calling variants in diploid systems		💻 ▾	📋			⚙ ▾
Calling variants in non-diploid systems prokaryote		💻 ▾	📋			
Exome sequencing data analysis		💻 ▾	📋	🔗	🪄	⚙ ▾
Mapping and molecular identification of phenotype-causing mutations		💻 ▾	📋	🔗	🪄	⚙ ▾

1. Hands-On Exome Variant Analysis

The data for the training:

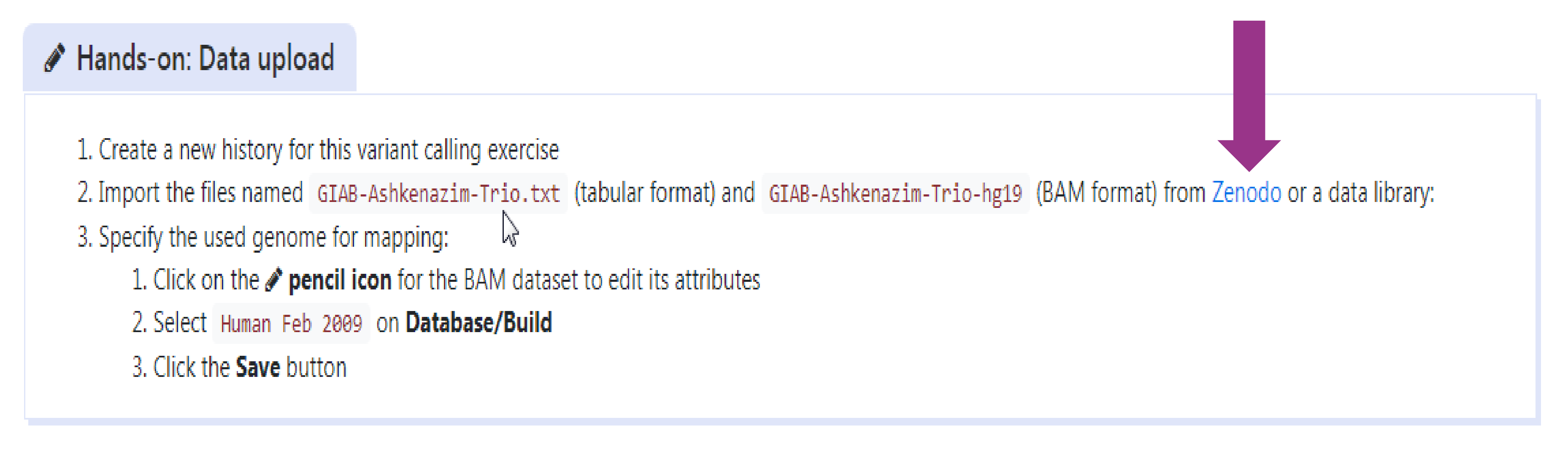
The Ashkenazim Father-Mother-Son trio

- HG002 - NA24385 - huAA53E0 (son)
- HG003 - NA24149 - hu6E4515 (father)
- HG004 - NA24143 - hu8E87A9 (mother)

Restricting alignments to a small portion of chromosome 19 containing the *POLRMT* gene

1. Hands-On Exome Variant Analysis

Download data from:

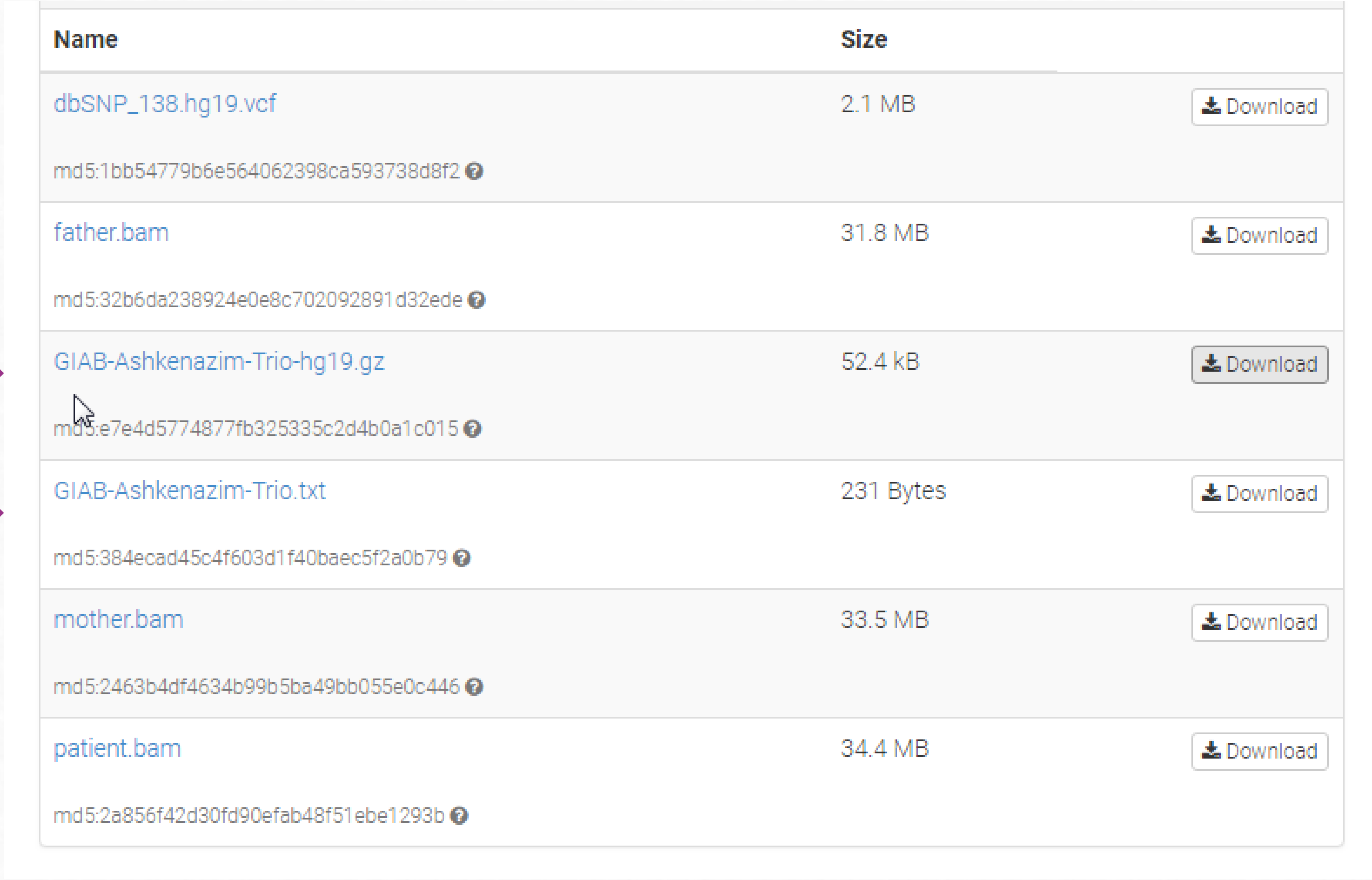


✏ Hands-on: Data upload

1. Create a new history for this variant calling exercise
2. Import the files named `GIAB-Ashkenazim-Trio.txt` (tabular format) and `GIAB-Ashkenazim-Trio-hg19` (BAM format) from Zenodo or a data library:
3. Specify the used genome for mapping:
    1. Click on the ✏ **pencil icon** for the BAM dataset to edit its attributes
    2. Select `Human Feb 2009` on **Database/Build**
    3. Click the **Save** button

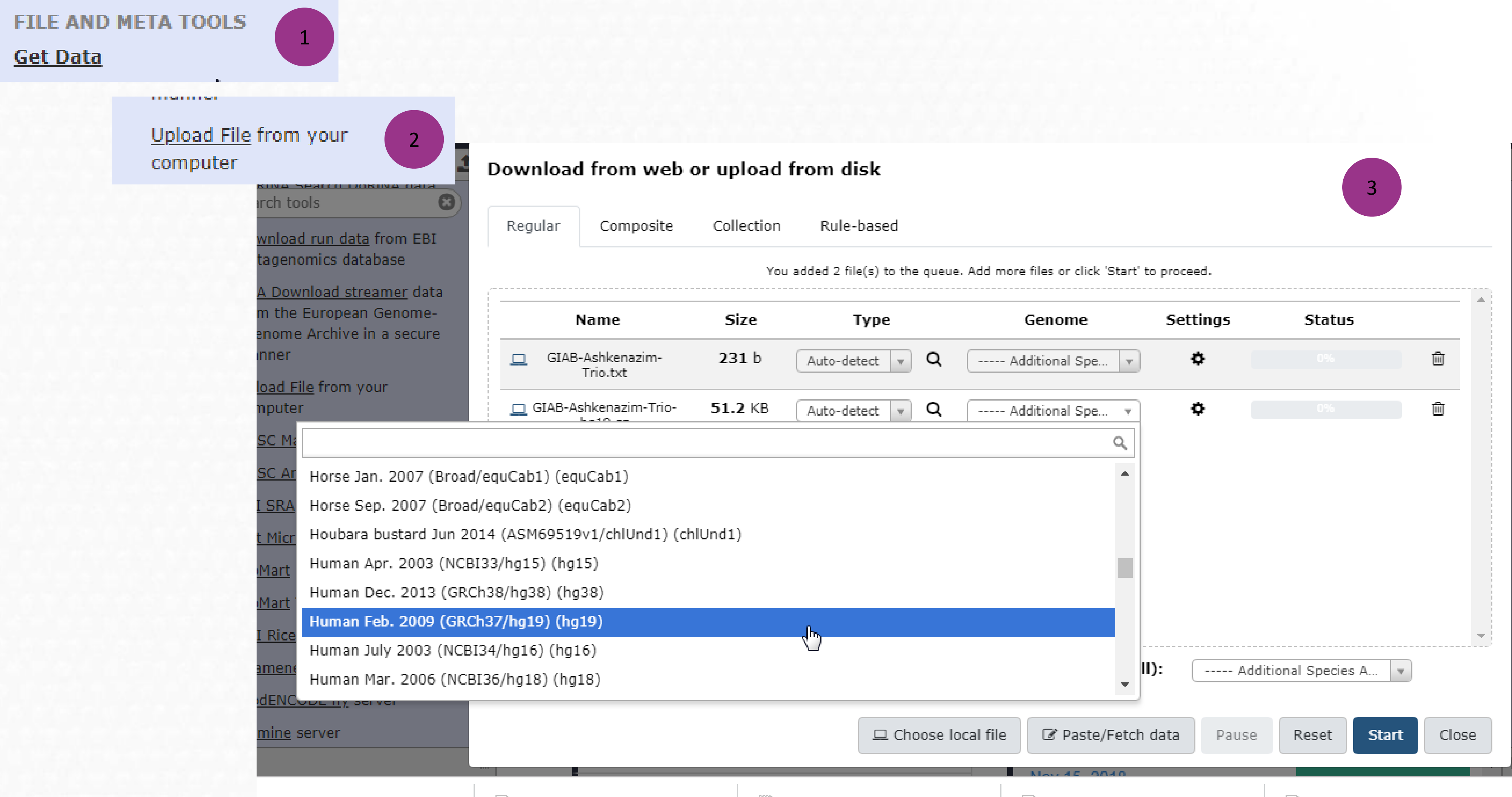
1. Hands-On Exome Variant Analysis

Download data from:



Name	Size	
dbSNP_138.hg19.vcf md5:1bb54779b6e564062398ca593738d8f2 ?	2.1 MB	Download
father.bam md5:32b6da238924e0e8c702092891d32ede ?	31.8 MB	Download
GIAB-Ashkenazim-Trio-hg19.gz md5:e7e4d5774877fb325335c2d4b0a1c015 ?	52.4 kB	Download
GIAB-Ashkenazim-Trio.txt md5:384ecad45c4f603d1f40baec5f2a0b79 ?	231 Bytes	Download
mother.bam md5:2463b4df4634b99b5ba49bb055e0c446 ?	33.5 MB	Download
patient.bam md5:2a856f42d30fd90efab48f51ebe1293b ?	34.4 MB	Download

Upload data to Galaxy



FILE AND META TOOLS

Get Data

Upload File from your computer

arch tools

wnload run data from EBI tagenomics database

A Download streamer data m the European Genome-enome Archive in a secure nner

load File from your mputer

mine server

Download from web or upload from disk

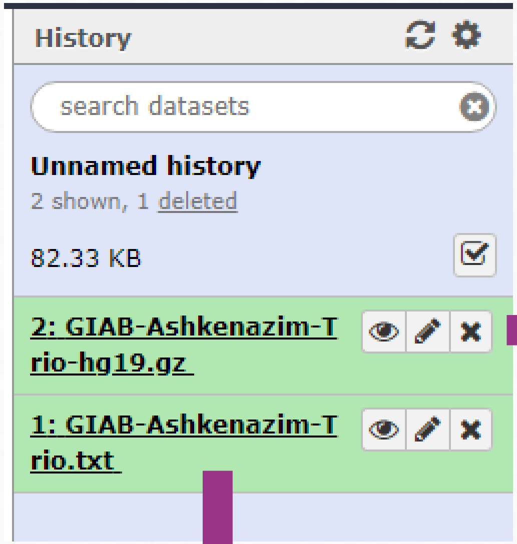
You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
GIAB-Ashkenazim-Trio.txt	231 b	Auto-detect	----- Additional Spe...		0%
GIAB-Ashkenazim-Trio-hg19...	51.2 KB	Auto-detect	----- Additional Spe...		0%

- Horse Jan. 2007 (Broad/equCab1) (equCab1)
- Horse Sep. 2007 (Broad/equCab2) (equCab2)
- Houbara bustard Jun 2014 (ASM69519v1/chlUnd1) (chlUnd1)
- Human Apr. 2003 (NCBI33/hg15) (hg15)
- Human Dec. 2013 (GRCh38/hg38) (hg38)
- **Human Feb. 2009 (GRCh37/hg19) (hg19)**
- Human July 2003 (NCBI34/hg16) (hg16)
- Human Mar. 2006 (NCBI36/hg18) (hg18)

ll): ----- Additional Species A...

1. Hands-On Exome Variant Analysis



History

search datasets

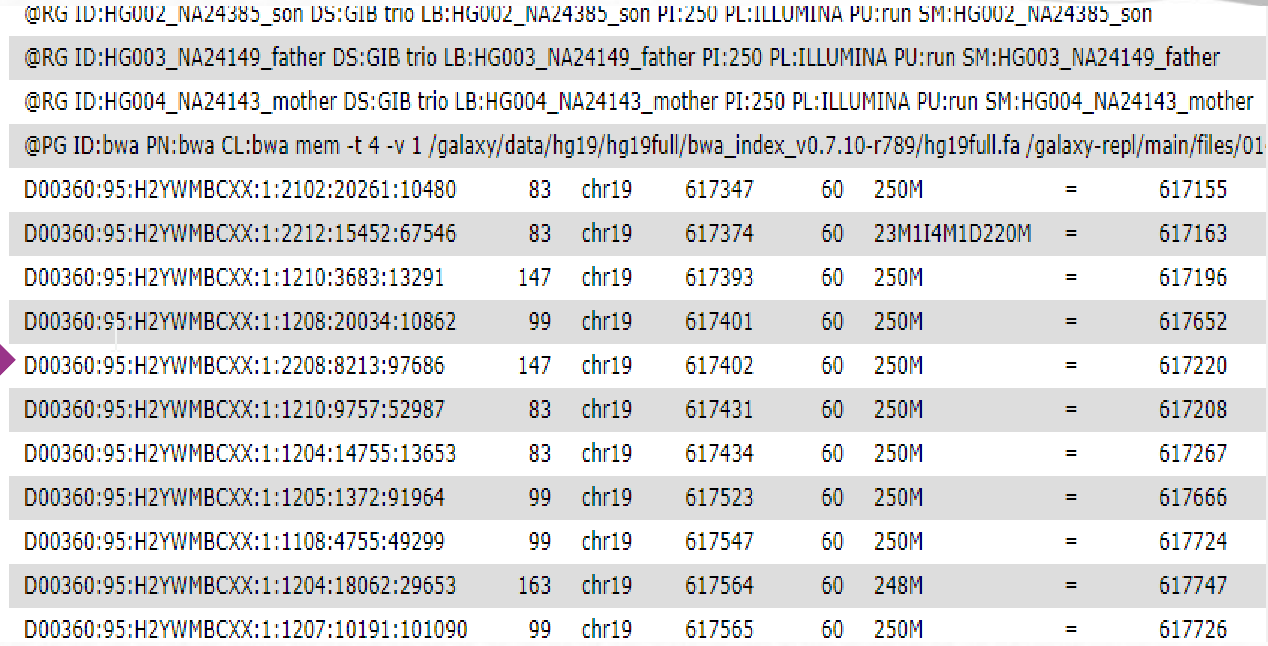
Unnamed history

2 shown, 1 deleted

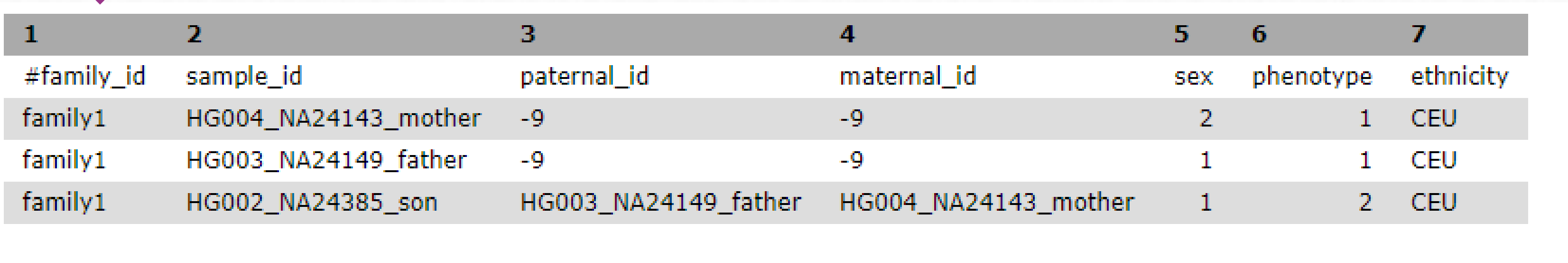
82.33 KB

2: GIAB-Ashkenazim-Trio-hg19.gz

1: GIAB-Ashkenazim-Trio.txt

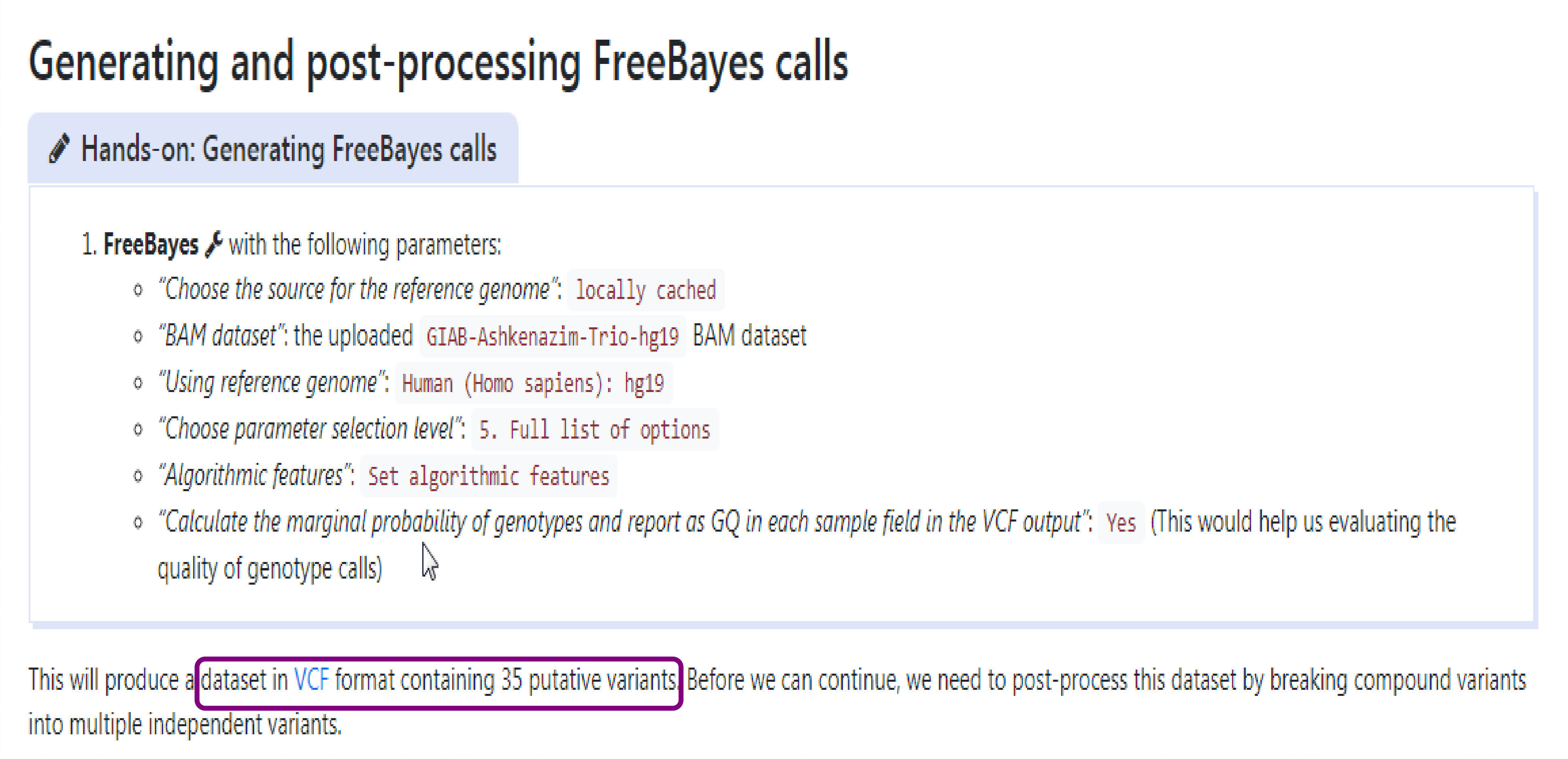


```
@RG ID:HG002_NA24385_son DS:GIB trio LB:HG002_NA24385_son PI:250 PL:ILLUMINA PU:run SM:HG002_NA24385_son
@RG ID:HG003_NA24149_father DS:GIB trio LB:HG003_NA24149_father PI:250 PL:ILLUMINA PU:run SM:HG003_NA24149_father
@RG ID:HG004_NA24143_mother DS:GIB trio LB:HG004_NA24143_mother PI:250 PL:ILLUMINA PU:run SM:HG004_NA24143_mother
@PG ID:bwa PN:bwa CL:bwa mem -t 4 -v 1 /galaxy/data/hg19/hg19full/bwa_index_v0.7.10-r789/hg19full.fa /galaxy-repl/main/files/01
D00360:95:H2YWMBCXX:1:2102:20261:10480     83   chr19   617347   60   250M          =   617155
D00360:95:H2YWMBCXX:1:2212:15452:67546     83   chr19   617374   60   23M1I4M1D220M =   617163
D00360:95:H2YWMBCXX:1:1210:3683:13291     147   chr19   617393   60   250M          =   617196
D00360:95:H2YWMBCXX:1:1208:20034:10862     99   chr19   617401   60   250M          =   617652
D00360:95:H2YWMBCXX:1:2208:8213:97686     147   chr19   617402   60   250M          =   617220
D00360:95:H2YWMBCXX:1:1210:9757:52987      83   chr19   617431   60   250M          =   617208
D00360:95:H2YWMBCXX:1:1204:14755:13653     83   chr19   617434   60   250M          =   617267
D00360:95:H2YWMBCXX:1:1205:1372:91964      99   chr19   617523   60   250M          =   617666
D00360:95:H2YWMBCXX:1:1108:4755:49299      99   chr19   617547   60   250M          =   617724
D00360:95:H2YWMBCXX:1:1204:18062:29653    163   chr19   617564   60   248M          =   617747
D00360:95:H2YWMBCXX:1:1207:10191:101090    99   chr19   617565   60   250M          =   617726
```



1	2	3	4	5	6	7
#family_id	sample_id	paternal_id	maternal_id	sex	phenotype	ethnicity
family1	HG004_NA24143_mother	-9	-9	2	1	CEU
family1	HG003_NA24149_father	-9	-9	1	1	CEU
family1	HG002_NA24385_son	HG003_NA24149_father	HG004_NA24143_mother	1	2	CEU

1. Hands-On Exome Variant Analysis



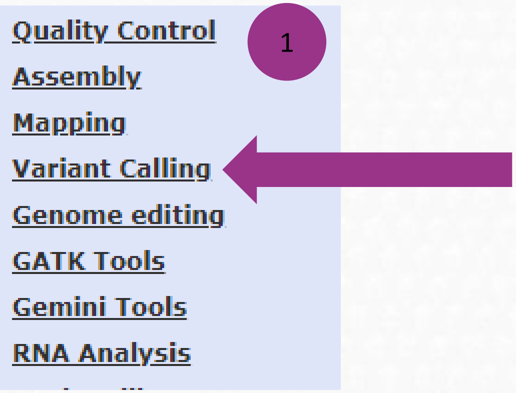
Generating and post-processing FreeBayes calls

Hands-on: Generating FreeBayes calls

1. **FreeBayes** with the following parameters:
   - *"Choose the source for the reference genome"*: `locally cached`
   - *"BAM dataset"*: the uploaded `GIAB-Ashkenazim-Trio-hg19` BAM dataset
   - *"Using reference genome"*: `Human (Homo sapiens): hg19`
   - *"Choose parameter selection level"*: `5. Full list of options`
   - *"Algorithmic features"*: `Set algorithmic features`
   - *"Calculate the marginal probability of genotypes and report as GQ in each sample field in the VCF output"*: `Yes` (This would help us evaluating the quality of genotype calls)

This will produce a dataset in VCF format containing 35 putative variants. Before we can continue, we need to post-process this dataset by breaking compound variants into multiple independent variants.

1. Hands-On Exome Variant Analysis



Quality Control

Assembly

Mapping

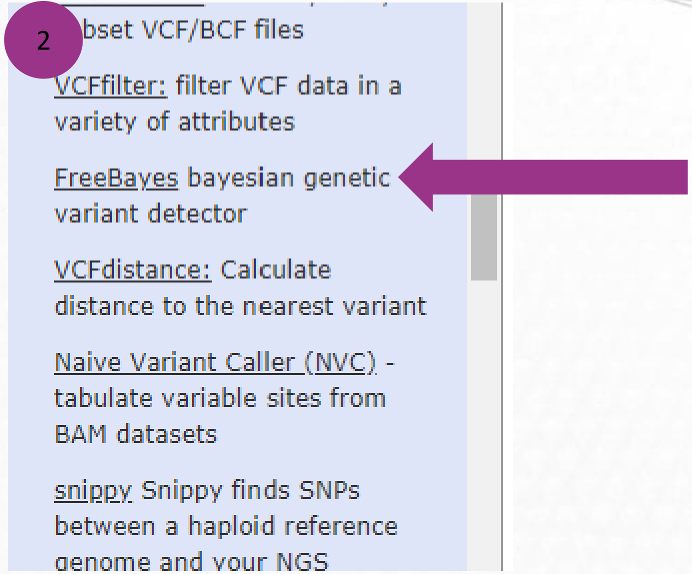
Variant Calling

Genome editing

GATK Tools

Gemini Tools

RNA Analysis



bset VCF/BCF files

VCFfilter: filter VCF data in a variety of attributes

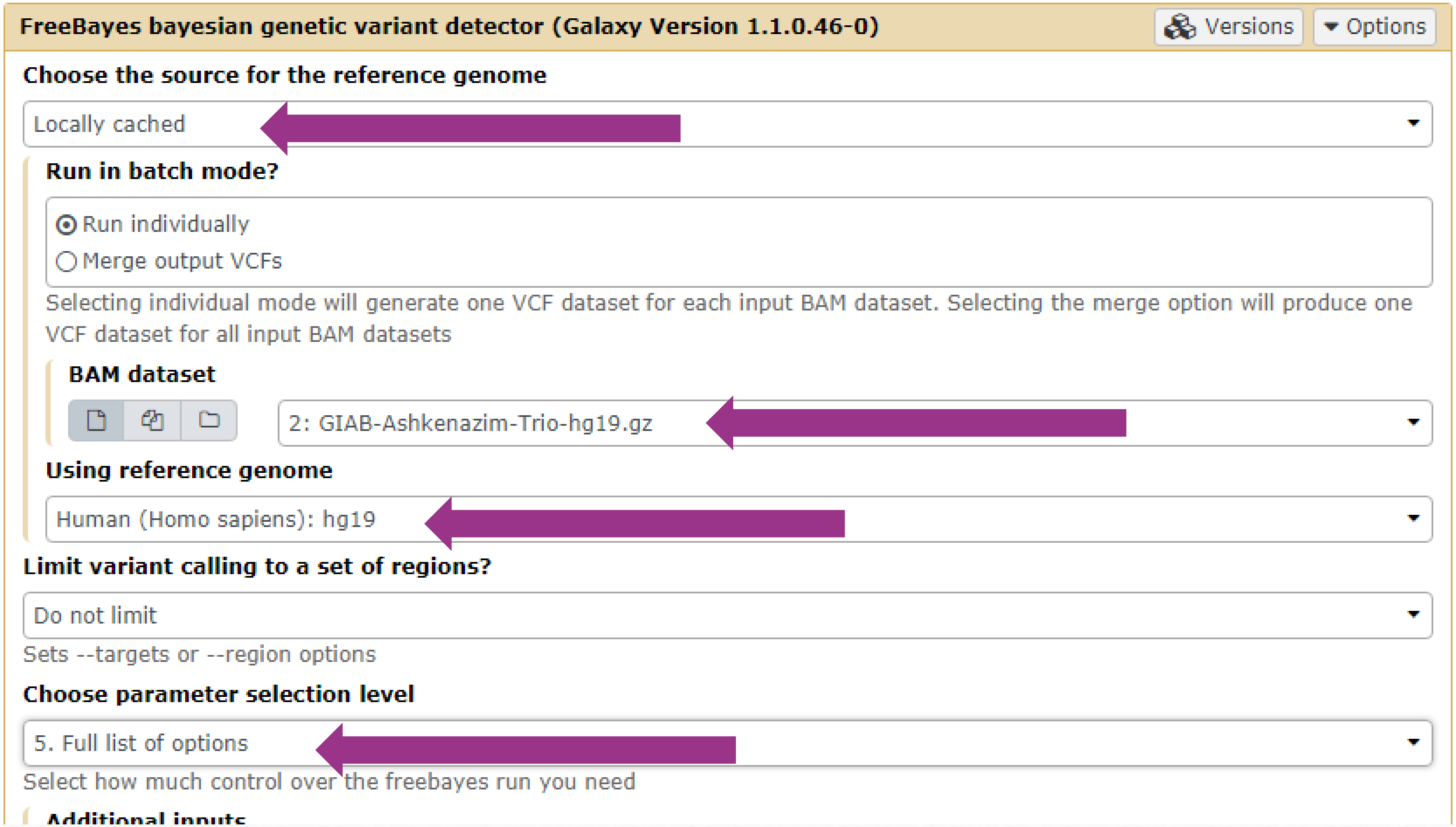
FreeBayes bayesian genetic variant detector

VCFdistance: Calculate distance to the nearest variant

Naive Variant Caller (NVC) - tabulate variable sites from BAM datasets

snippy Snippy finds SNPs between a haploid reference genome and your NGS

1. Hands-On Exome Variant Analysis



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

2: GIAB-Ashkenazim-Trio-hg19.gz

Using reference genome

Human (Homo sapiens): hg19

Limit variant calling to a set of regions?

Do not limit

Sets --targets or --region options

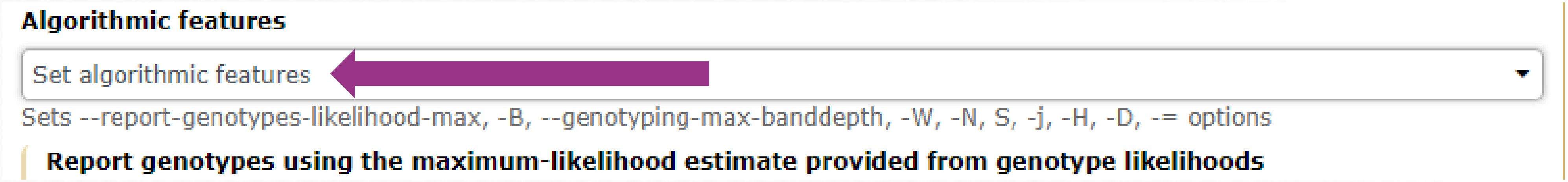
Choose parameter selection level

5. Full list of options

Select how much control over the freebayes run you need

Additional inputs

1. Hands-On Exome Variant Analysis

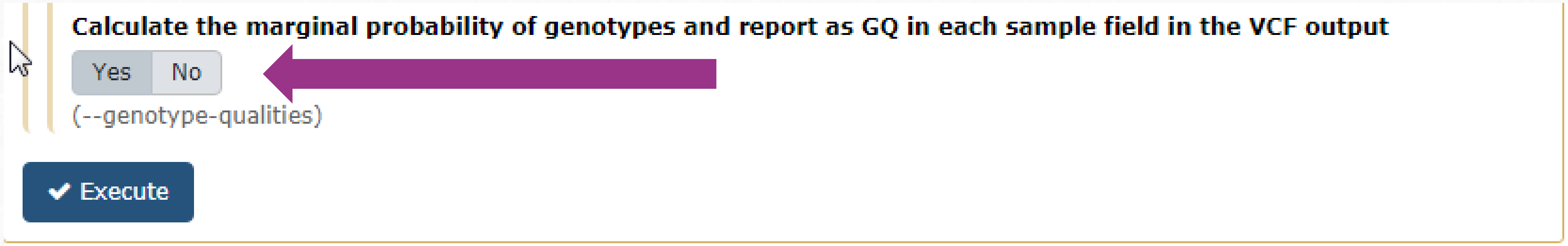


Algorithmic features

Set algorithmic features

Sets --report-genotypes-likelihood-max, -B, --genotyping-max-banddepth, -W, -N, S, -j, -H, -D, -= options

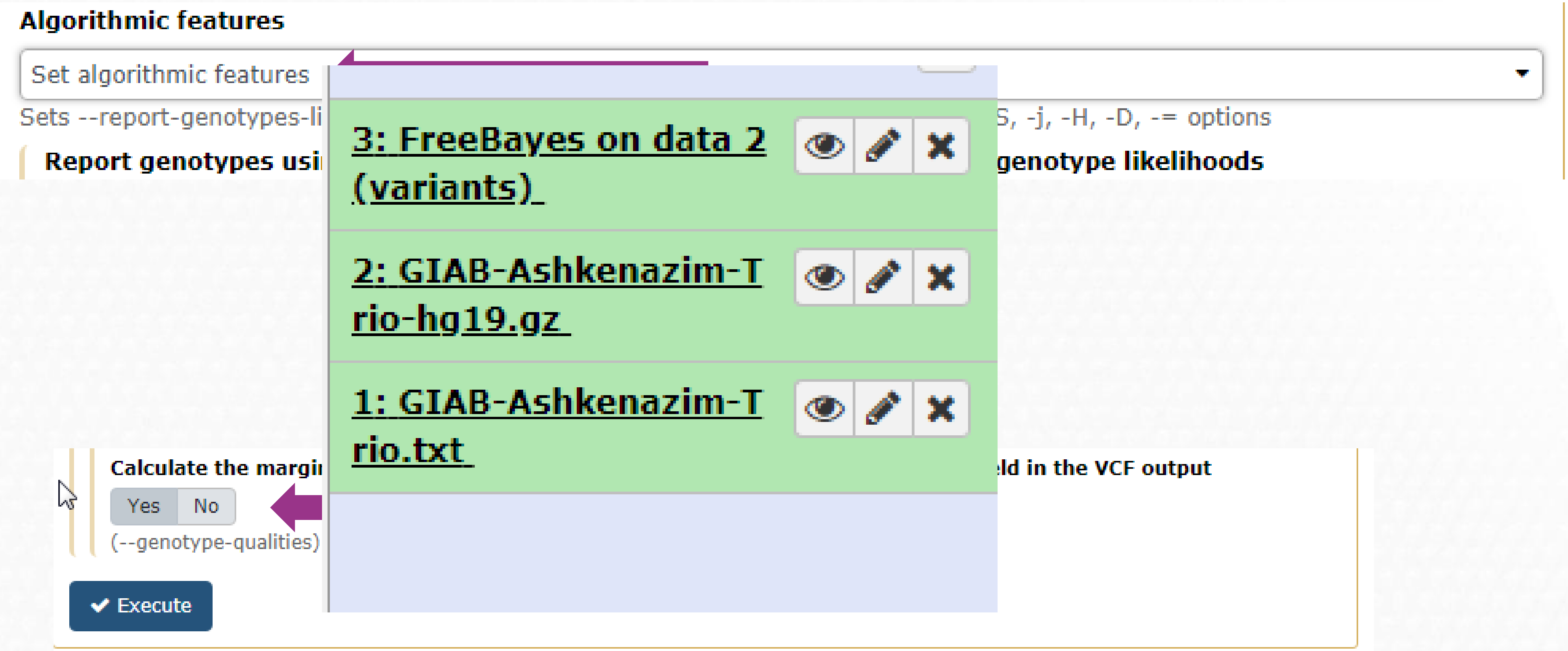
Report genotypes using the maximum-likelihood estimate provided from genotype likelihoods



Calculate the marginal probability of genotypes and report as GQ in each sample field in the VCF output

(--genotype-qualities)

1. Hands-On Exome Variant Analysis



Algorithmic features

Set algorithmic features

Sets --report-genotypes-li S, -j, -H, -D, -= options

Report genotypes usi genotype likelihoods

3: FreeBayes on data 2 (variants)

2: GIAB-Ashkenazim-Trio-hg19.gz

1: GIAB-Ashkenazim-Trio.txt

Calculate the margi eld in the VCF output

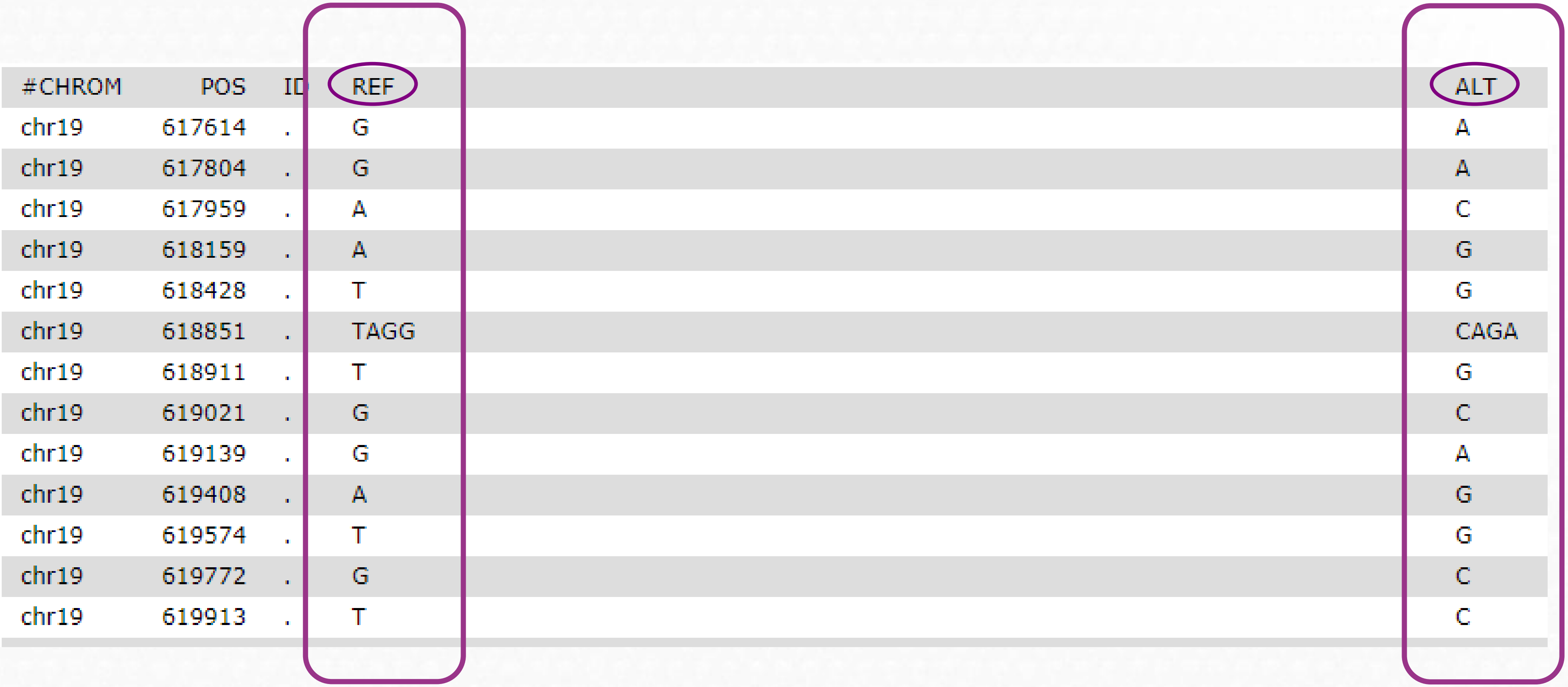
Yes No

(--genotype-qualities)

✓ Execute

1. Hands-On Exome Variant Analysis

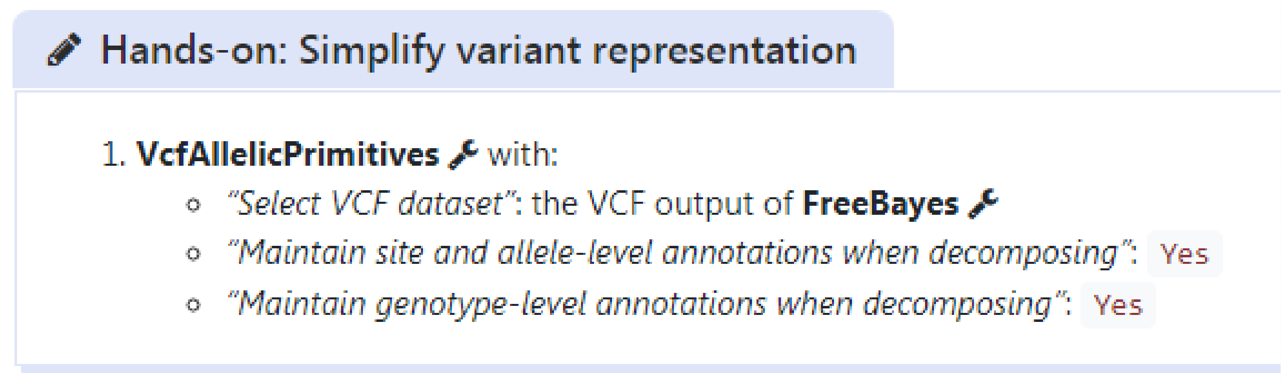
This will produce a dataset in **VCF** format containing **35 putative variants**.



#CHROM	POS	ID	REF	ALT
chr19	617614	.	G	A
chr19	617804	.	G	A
chr19	617959	.	A	C
chr19	618159	.	A	G
chr19	618428	.	T	G
chr19	618851	.	TAGG	CAGA
chr19	618911	.	T	G
chr19	619021	.	G	C
chr19	619139	.	G	A
chr19	619408	.	A	G
chr19	619574	.	T	G
chr19	619772	.	G	C
chr19	619913	.	T	C

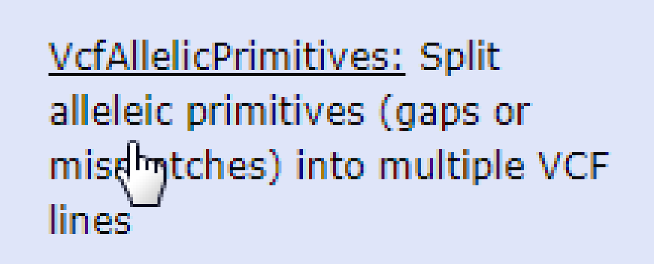
1. Hands-On Exome Variant Analysis

Before we can continue, we need to post-process this dataset by **breaking compound variants into multiple independent variants**.



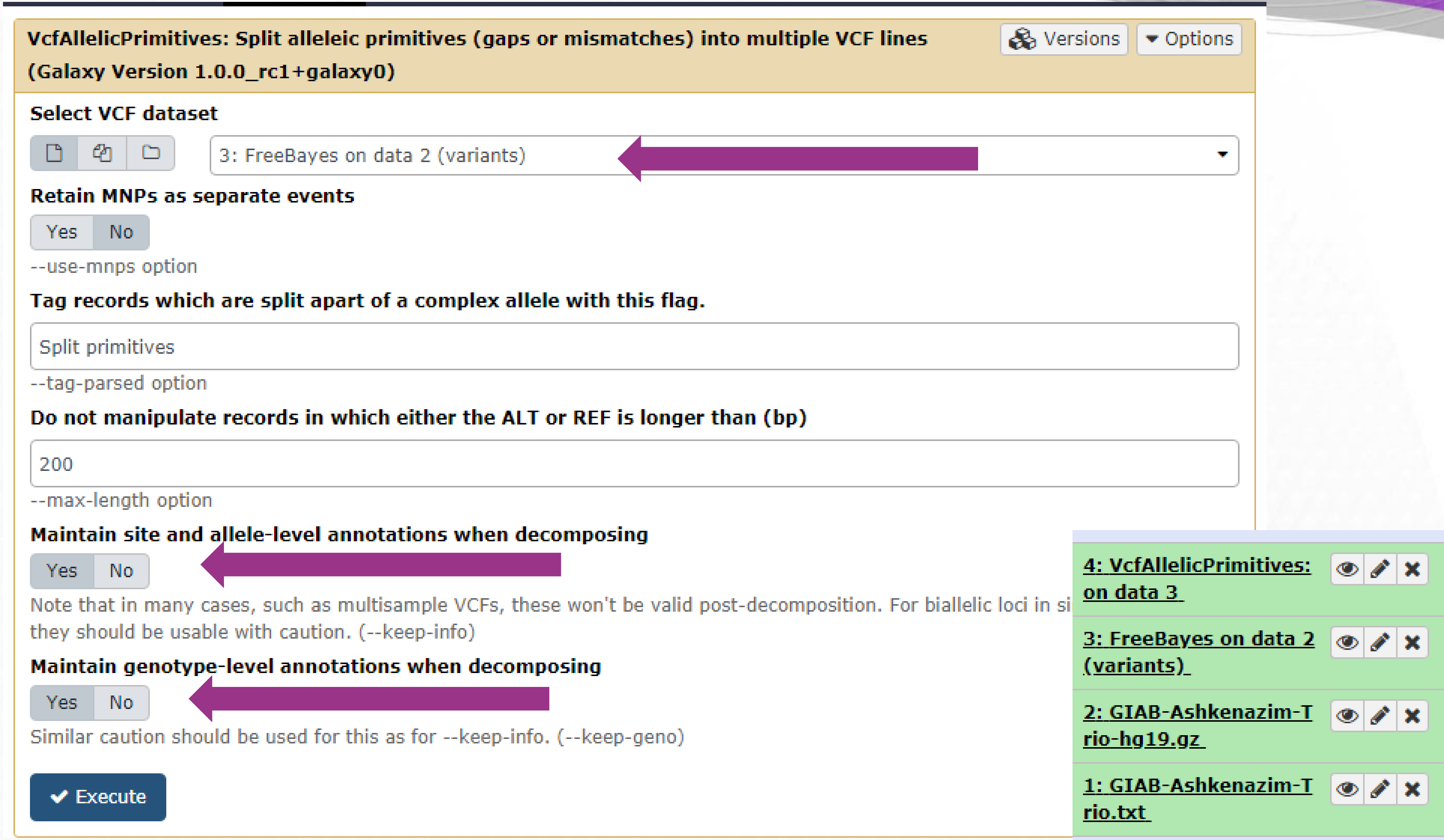
Hands-on: Simplify variant representation

1. **VcfAllelicPrimitives** with:
   - *"Select VCF dataset"*: the VCF output of **FreeBayes**
   - *"Maintain site and allele-level annotations when decomposing"*: `Yes`
   - *"Maintain genotype-level annotations when decomposing"*: `Yes`



VcfAllelicPrimitives: Split alleleic primitives (gaps or mismatches) into multiple VCF lines

1. Hands-On Exome Variant Analysis



VcfAllelicPrimitives: Split alleleic primitives (gaps or mismatches) into multiple VCF lines (Galaxy Version 1.0.0_rc1+galaxy0)

Select VCF dataset

3: FreeBayes on data 2 (variants)

Retain MNPs as separate events

Yes No

--use-mnps option

Tag records which are split apart of a complex allele with this flag.

Split primitives

--tag-parsed option

Do not manipulate records in which either the ALT or REF is longer than (bp)

200

--max-length option

Maintain site and allele-level annotations when decomposing

Yes No

Note that in many cases, such as multisample VCFs, these won't be valid post-decomposition. For biallelic loci in si they should be usable with caution. (--keep-info)

Maintain genotype-level annotations when decomposing

Yes No

Similar caution should be used for this as for --keep-info. (--keep-geno)

4: VcfAllelicPrimitives: on data 3

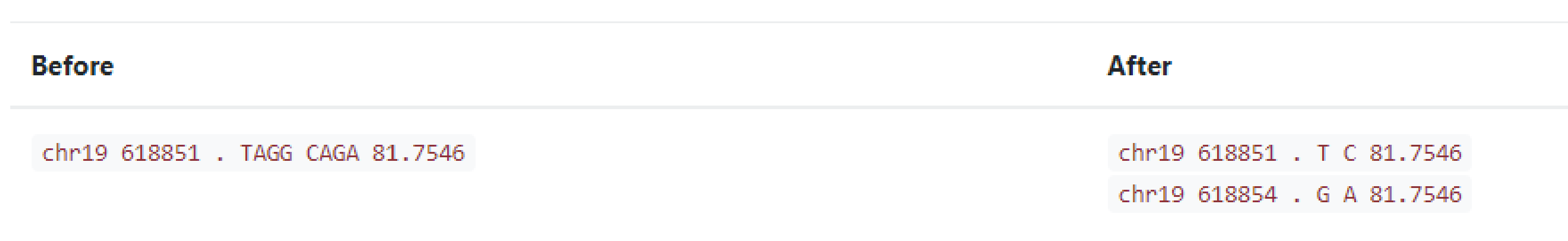
3: FreeBayes on data 2 (variants)

2: GIAB-Ashkenazim-Trio-hg19.gz

1: GIAB-Ashkenazim-Trio.txt

1. Hands-On Exome Variant Analysis

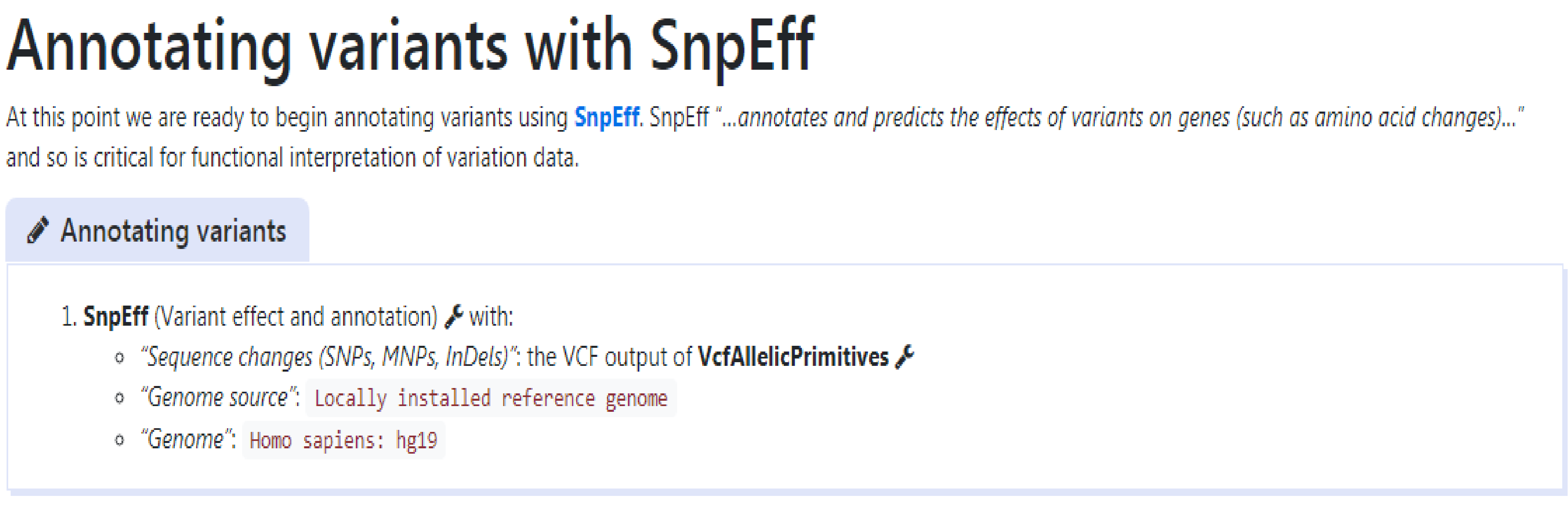
VCFAllelicPrimitives generates a VCF files containing **37 records** (the input VCF only contained **35**). This is because a multiple nucleotide polymorphism (TAGG|CAGA) at position 618851 have been converted to two.



Before	After

`chr19 618851 . TAGG CAGA 81.7546`	`chr19 618851 . T C 81.7546`
	`chr19 618854 . G A 81.7546`

1. Hands-On Exome Variant Analysis

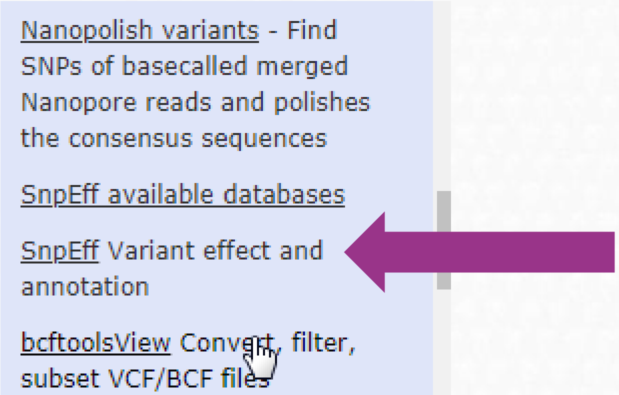


Annotating variants with SnpEff

At this point we are ready to begin annotating variants using **SnpEff**. SnpEff "*...annotates and predicts the effects of variants on genes (such as amino acid changes)...*" and so is critical for functional interpretation of variation data.

Annotating variants

1. **SnpEff** (Variant effect and annotation) with:
   - *"Sequence changes (SNPs, MNPs, InDels)"*: the VCF output of **VcfAllelicPrimitives**
   - *"Genome source"*: `Locally installed reference genome`
   - *"Genome"*: `Homo sapiens: hg19`



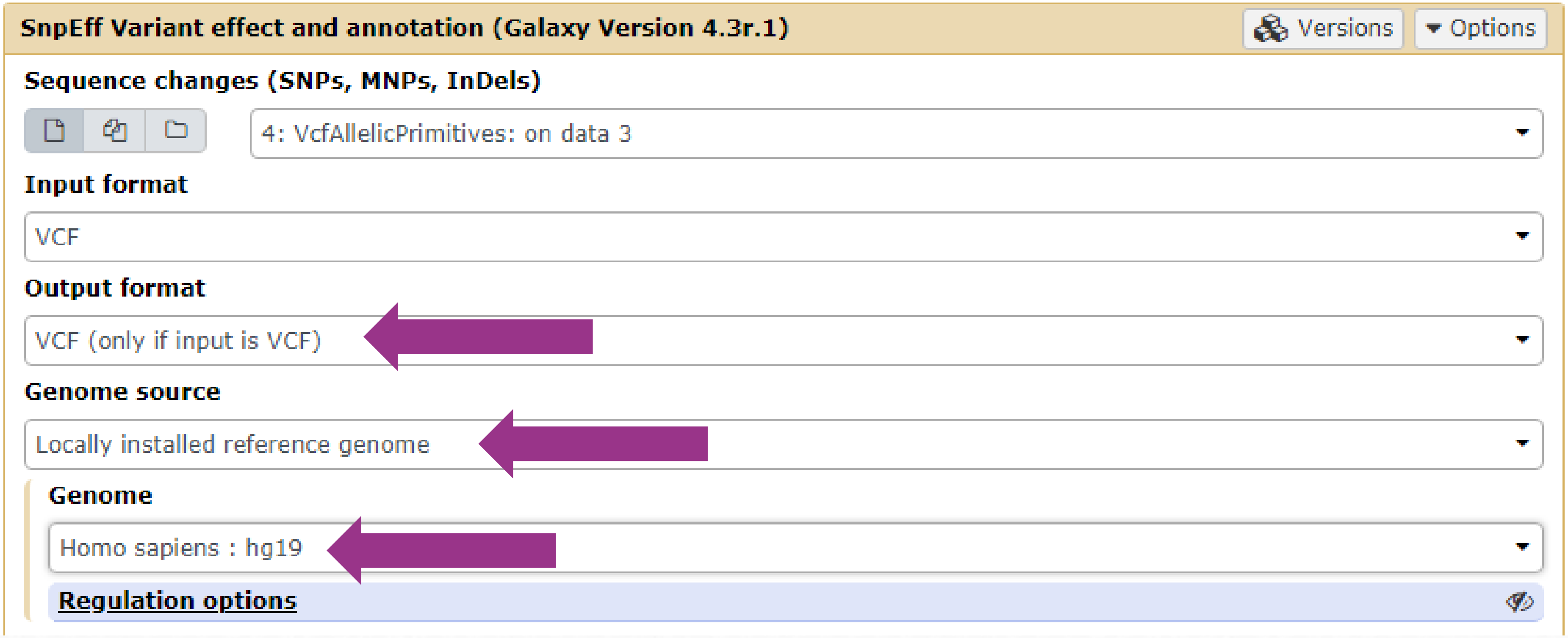
Nanopolish variants - Find SNPs of basecalled merged Nanopore reads and polishes the consensus sequences

SnpEff available databases

SnpEff Variant effect and annotation

bcftoolsView Convert, filter, subset VCF/BCF files

1. Hands-On Exome Variant Analysis



SnpEff Variant effect and annotation (Galaxy Version 4.3r.1)

Sequence changes (SNPs, MNPs, InDels): 4: VcfAllelicPrimitives: on data 3

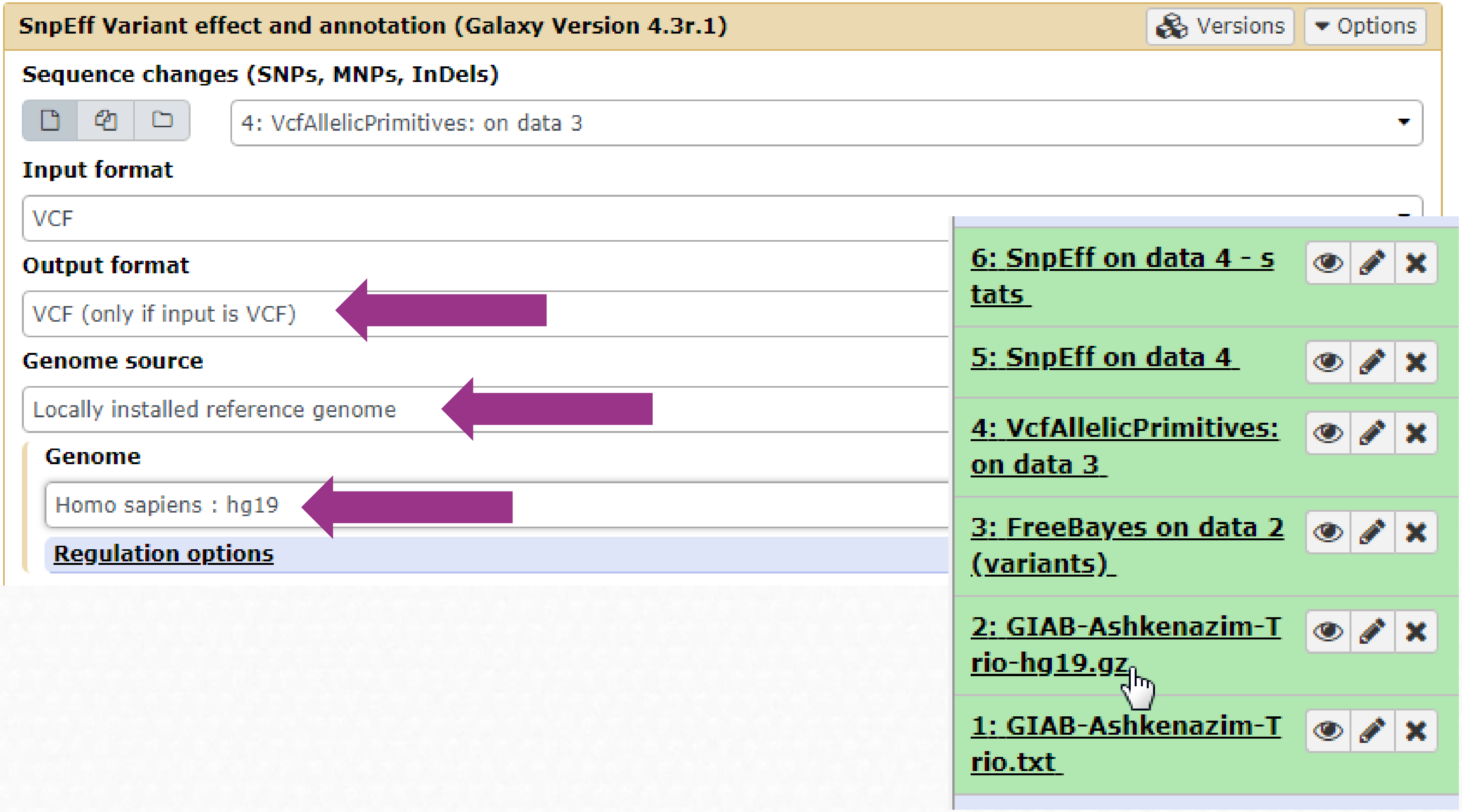
Input format: VCF

Output format: VCF (only if input is VCF)

Genome source: Locally installed reference genome

Genome: Homo sapiens : hg19

1. Hands-On Exome Variant Analysis



SnpEff Variant effect and annotation (Galaxy Version 4.3r.1)

Sequence changes (SNPs, MNPs, InDels): 4: VcfAllelicPrimitives: on data 3

Input format: VCF

Output format: VCF (only if input is VCF)

Genome source: Locally installed reference genome

Genome: Homo sapiens : hg19

Regulation options

6: SnpEff on data 4 - stats

5: SnpEff on data 4

4: VcfAllelicPrimitives: on data 3

3: FreeBayes on data 2 (variants)

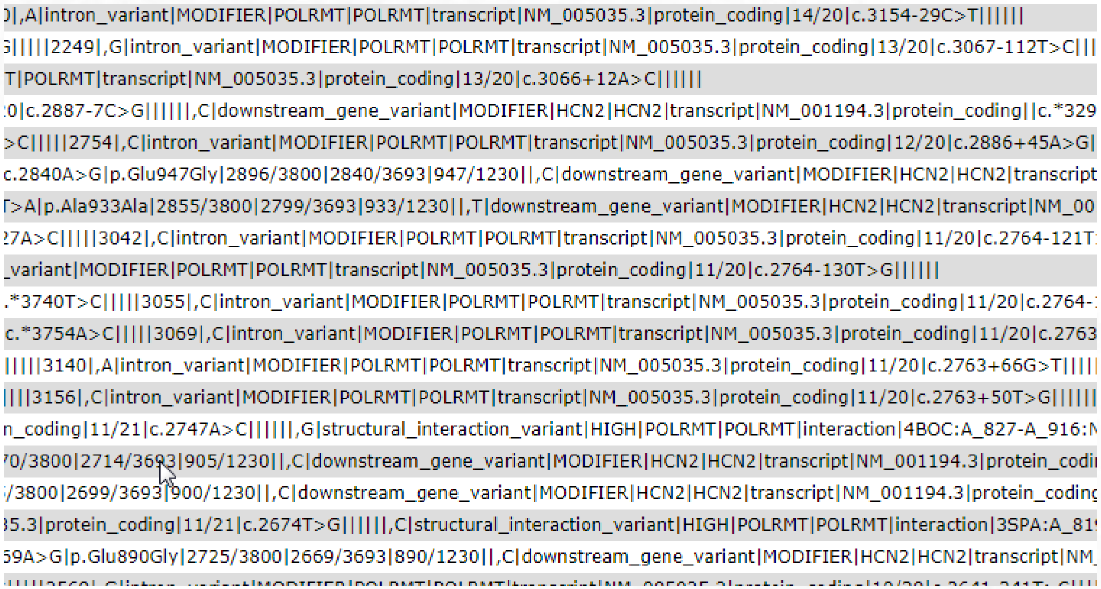
2: GIAB-Ashkenazim-Trio-hg19.gz

1: GIAB-Ashkenazim-Trio.txt

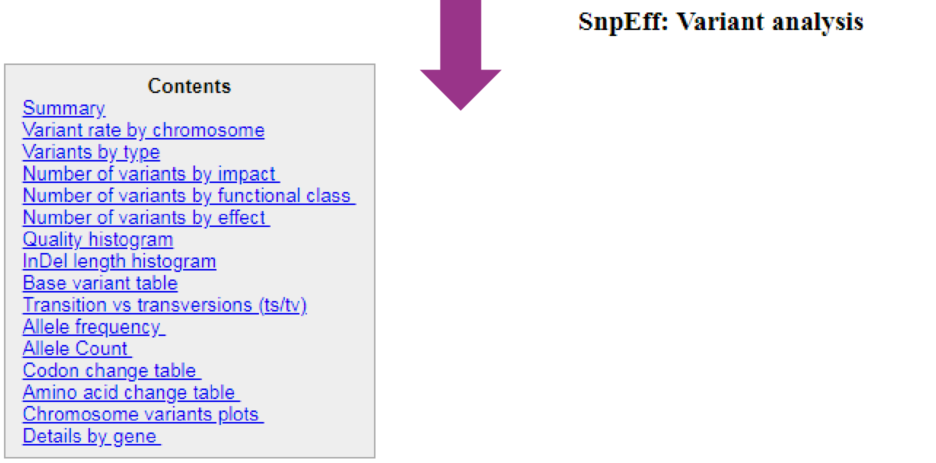
1. Hands-On Exome Variant Analysis

SnpEff will generate two outputs:

- an annotated VCF file
- an HTML report



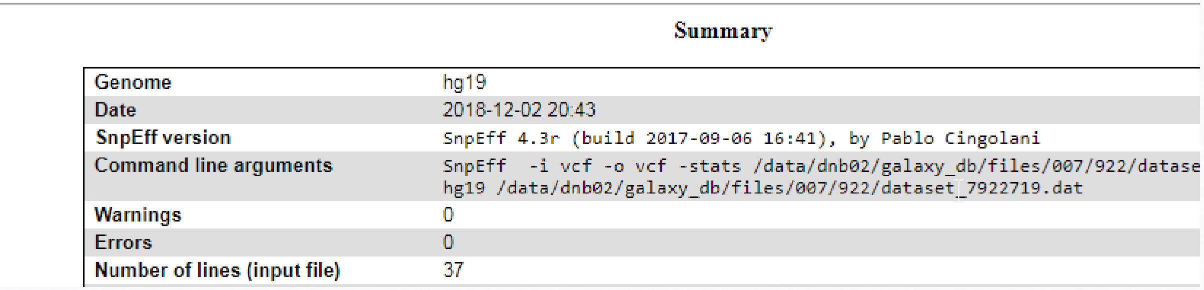
```
|,A|intron_variant|MODIFIER|POLRMT|POLRMT|transcript|NM_005035.3|protein_coding|14/20|c.3154-29C>T||||||
|||||2249|,G|intron_variant|MODIFIER|POLRMT|POLRMT|transcript|NM_005035.3|protein_coding|13/20|c.3067-112T>C|||
T|POLRMT|transcript|NM_005035.3|protein_coding|13/20|c.3066+12A>C||||||
0|c.2887-7C>G||||||,C|downstream_gene_variant|MODIFIER|HCN2|HCN2|transcript|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
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||||||
.*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|4BOC:A_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
35.3|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_
```



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
- Allele Count
- Codon change table
- Amino acid change table
- Chromosome variants plots
- Details by gene



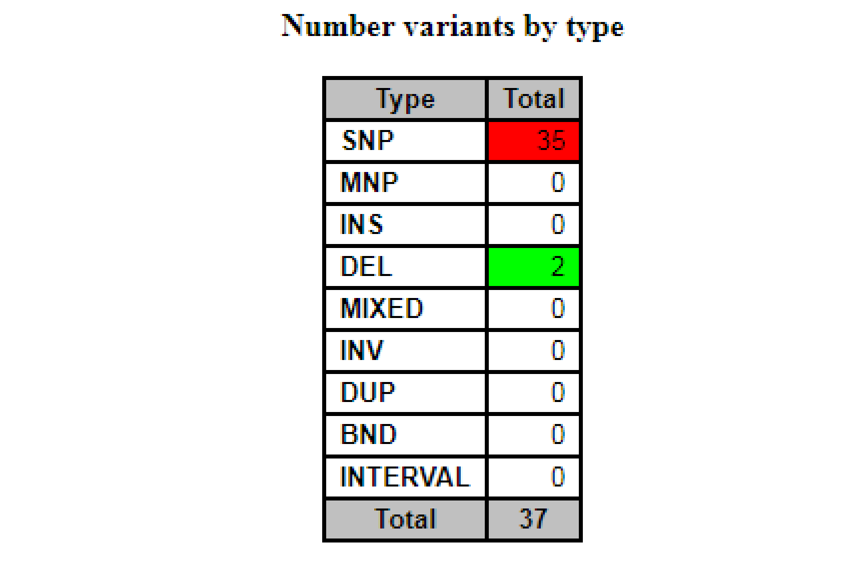
Summary

Genome	hg19
Date	2018-12-02 20:43
SnpEff version	`SnpEff 4.3r (build 2017-09-06 16:41), by Pablo Cingolani`

Command line arguments	`SnpEff -i vcf -o vcf -stats /data/dnb02/galaxy_db/files/007/922/datase`
	`hg19 /data/dnb02/galaxy_db/files/007/922/dataset_7922719.dat`

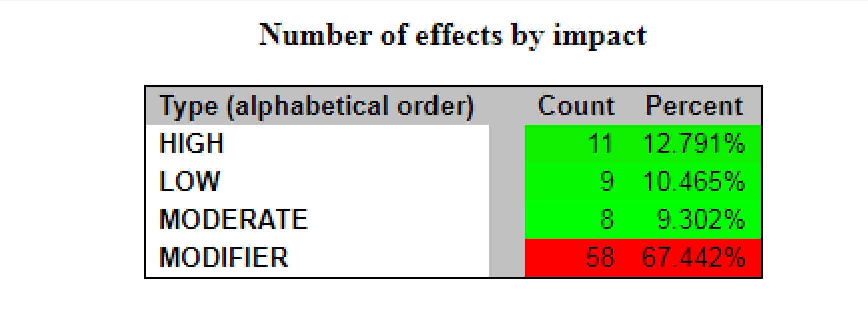
Warnings	0
Errors	0
Number of lines (input file)	37

1. Hands-On Exome Variant Analysis



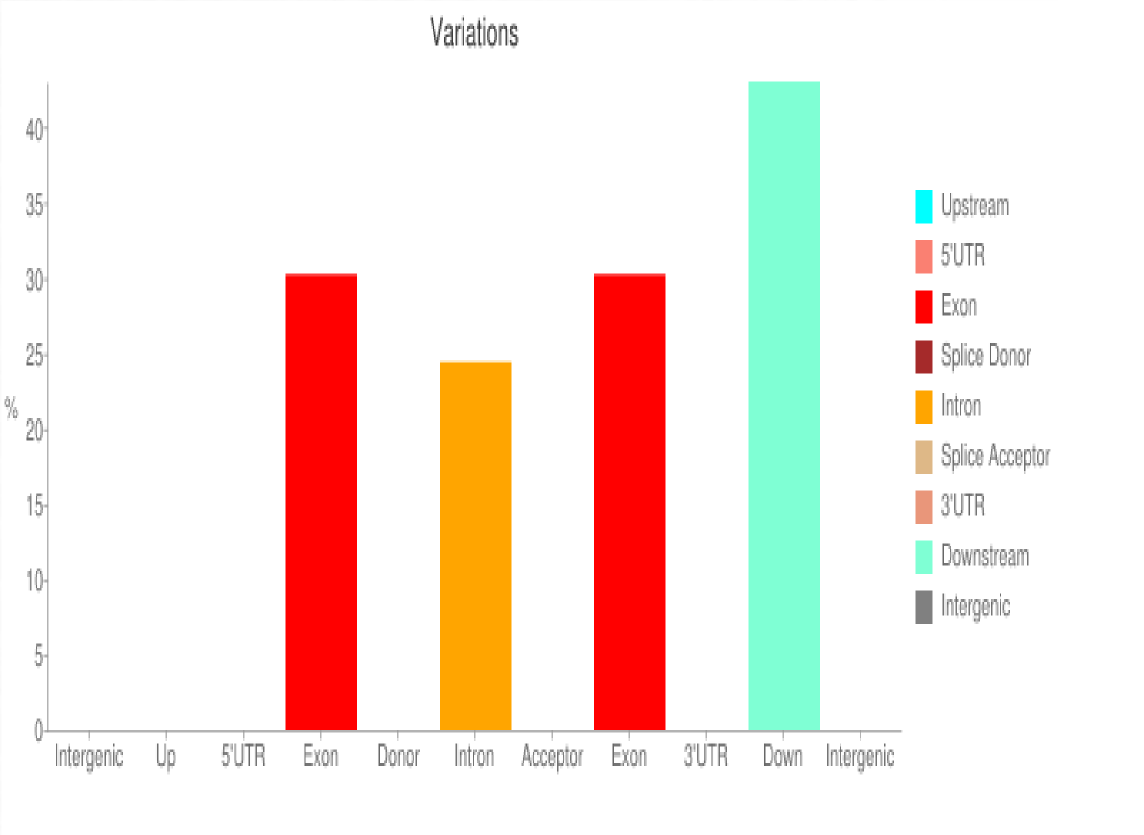
Number variants by type

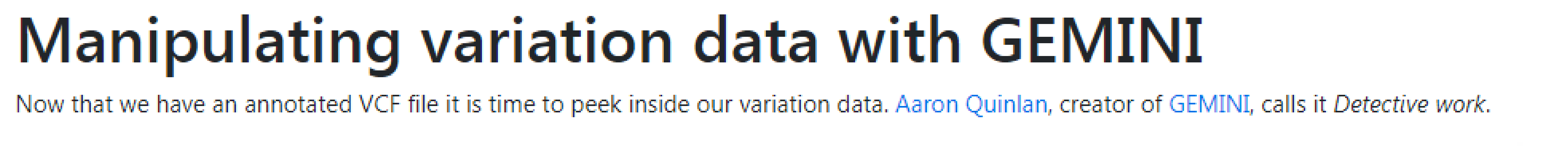
Type	Total
SNP	35
MNP	0
INS	0
DEL	2
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	37



Number of effects by impact

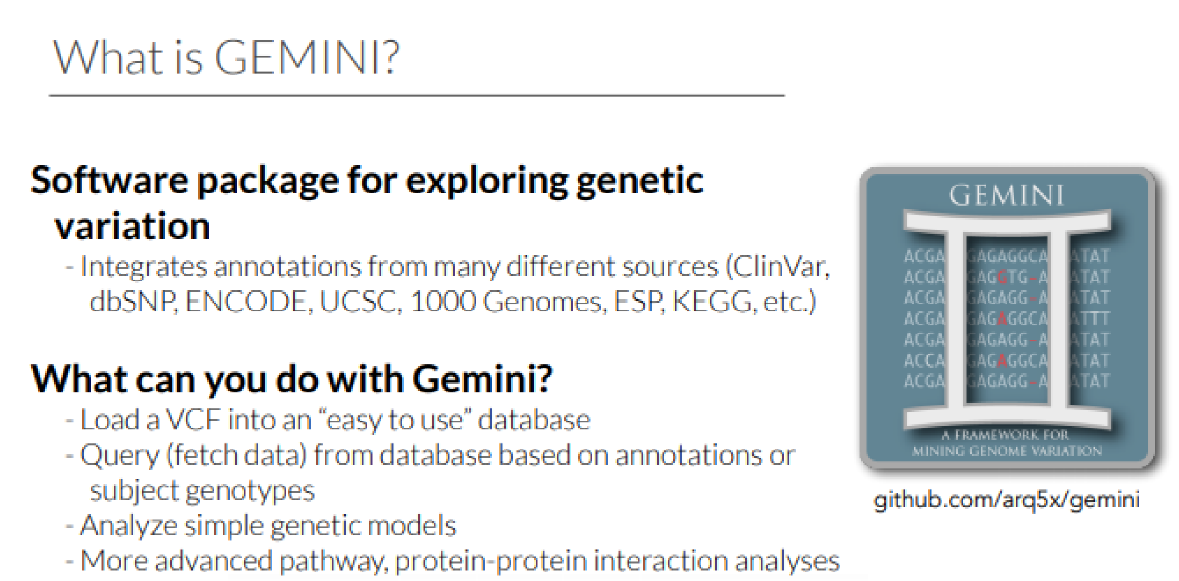
Type (alphabetical order)	Count	Percent
HIGH	11	12.791%
LOW	9	10.465%
MODERATE	8	9.302%
MODIFIER	58	67.442%





Manipulating variation data with GEMINI

Now that we have an annotated VCF file it is time to peek inside our variation data. Aaron Quinlan, creator of GEMINI, calls it *Detective work*.



What is GEMINI?

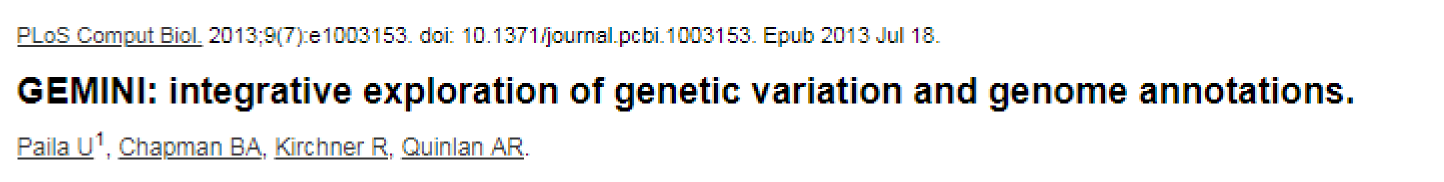
Software package for exploring genetic variation

- Integrates annotations from many different sources (ClinVar, dbSNP, ENCODE, UCSC, 1000 Genomes, ESP, KEGG, etc.)

What can you do with Gemini?

- Load a VCF into an “easy to use” database
- Query (fetch data) from database based on annotations or subject genotypes
- Analyze simple genetic models
- More advanced pathway, protein-protein interaction analyses

github.com/arq5x/gemini

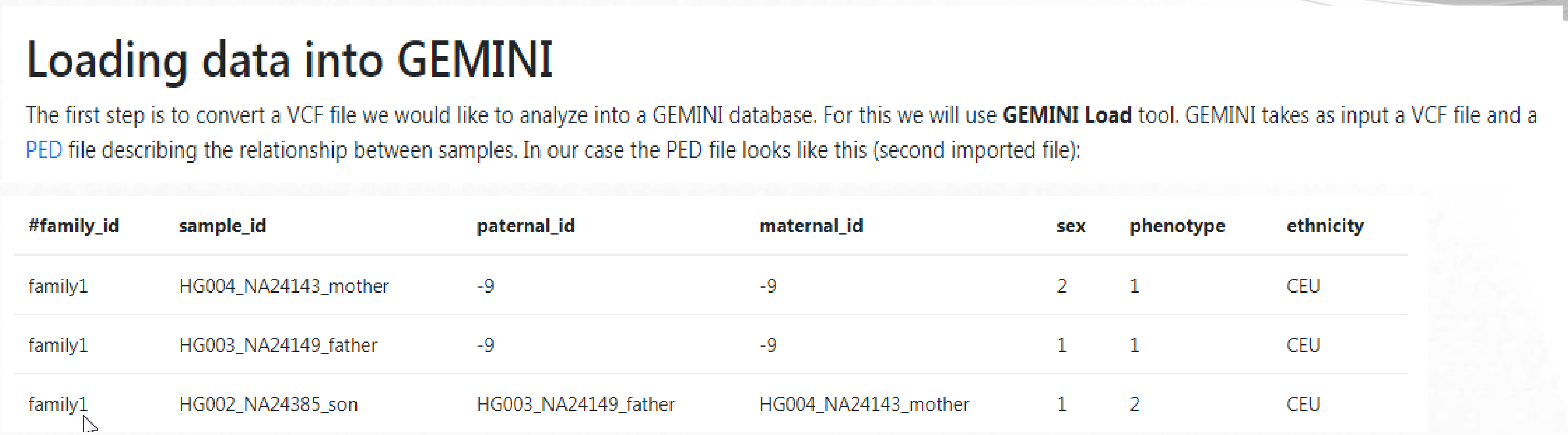


PLoS Comput Biol. 2013;9(7):e1003153. doi: 10.1371/journal.pcbi.1003153. Epub 2013 Jul 18.

GEMINI: integrative exploration of genetic variation and genome annotations.

Paila U[1], Chapman BA, Kirchner R, Quinlan AR.

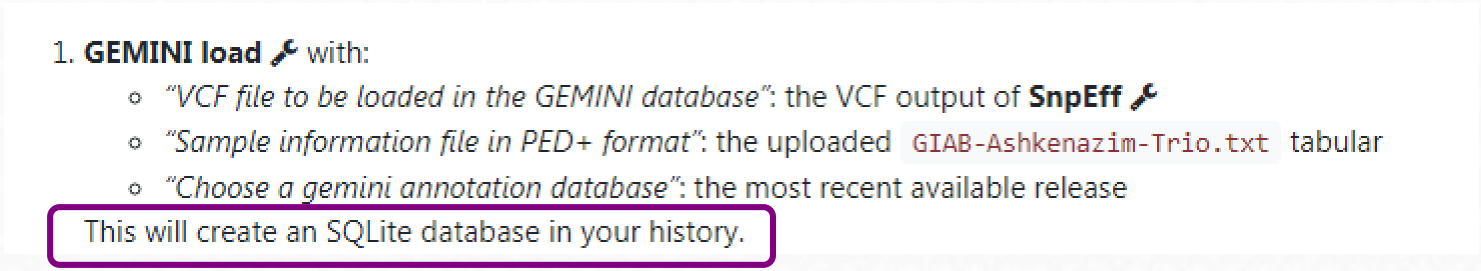
1. Hands-On Exome Variant Analysis



Loading data into GEMINI

The first step is to convert a VCF file we would like to analyze into a GEMINI database. For this we will use **GEMINI Load** tool. GEMINI takes as input a VCF file and a PED file describing the relationship between samples. In our case the PED file looks like this (second imported file):

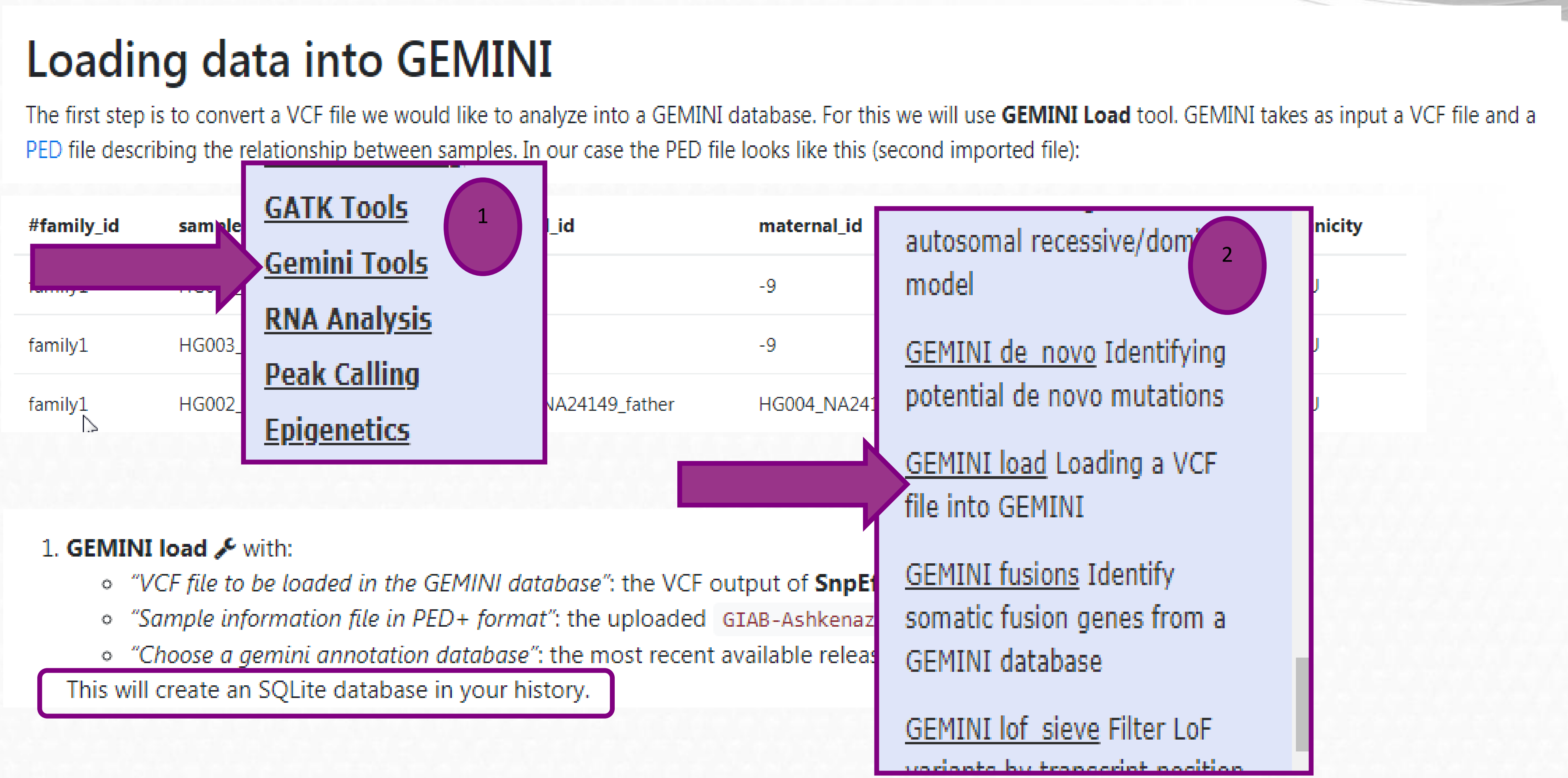
#family_id	sample_id	paternal_id	maternal_id	sex	phenotype	ethnicity
family1	HG004_NA24143_mother	-9	-9	2	1	CEU
family1	HG003_NA24149_father	-9	-9	1	1	CEU
family1	HG002_NA24385_son	HG003_NA24149_father	HG004_NA24143_mother	1	2	CEU



1. **GEMINI load** 🔧 with:
   - *"VCF file to be loaded in the GEMINI database"*: the VCF output of **SnpEff** 🔧
   - *"Sample information file in PED+ format"*: the uploaded `GIAB-Ashkenazim-Trio.txt` tabular
   - *"Choose a gemini annotation database"*: the most recent available release

This will create an SQLite database in your history.

1. Hands-On Exome Variant Analysis



Loading data into GEMINI

The first step is to convert a VCF file we would like to analyze into a GEMINI database. For this we will use **GEMINI Load** tool. GEMINI takes as input a VCF file and a PED file describing the relationship between samples. In our case the PED file looks like this (second imported file):

#family_id	sample	_id	maternal_id		nicity
			-9		U
family1	HG003_		-9		U
family1	HG002_	NA24149_father	HG004_NA241		U

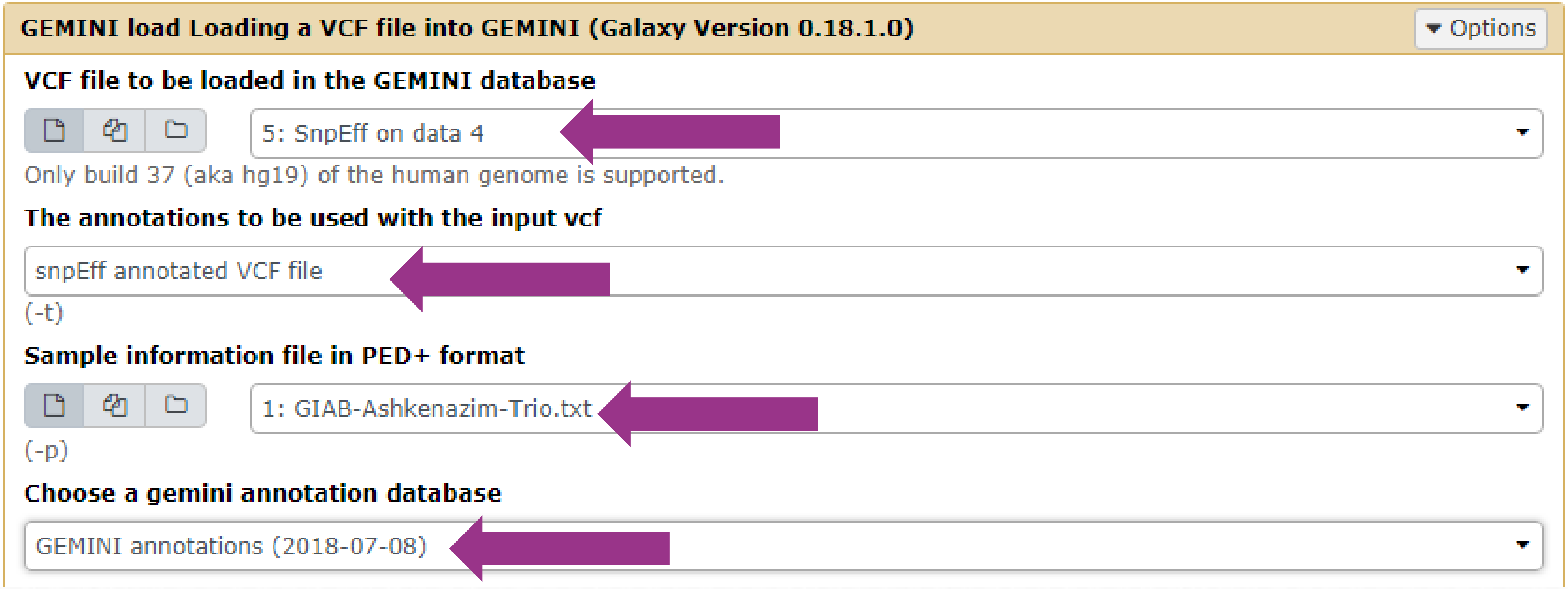
GATK Tools
Gemini Tools
RNA Analysis
Peak Calling
Epigenetics
1
autosomal recessive/dom
model
2
GEMINI de_novo Identifying potential de novo mutations
GEMINI load Loading a VCF file into GEMINI
GEMINI fusions Identify somatic fusion genes from a GEMINI database
GEMINI lof_sieve Filter LoF variants by transcript position

1. **GEMINI load** with:
   - *"VCF file to be loaded in the GEMINI database"*: the VCF output of **SnpE**
   - *"Sample information file in PED+ format"*: the uploaded `GIAB-Ashkenaz`
   - *"Choose a gemini annotation database"*: the most recent available releas

This will create an SQLite database in your history.

Loading data into GEMINI

The first step is to convert a VCF file we would like to analyze into a GEMINI database. For this we will use **GEMINI Load** tool. GEMINI takes as input a VCF file and a PED file describing the relationship between samples. In our case the PED file looks like this (second imported file):



GEMINI load Loading a VCF file into GEMINI (Galaxy Version 0.18.1.0)

VCF file to be loaded in the GEMINI database

5: SnpEff on data 4

Only build 37 (aka hg19) of the human genome is supported.

The annotations to be used with the input vcf

snpEff annotated VCF file

(-t)

Sample information file in PED+ format

1: GIAB-Ashkenazim-Trio.txt

(-p)

Choose a gemini annotation database

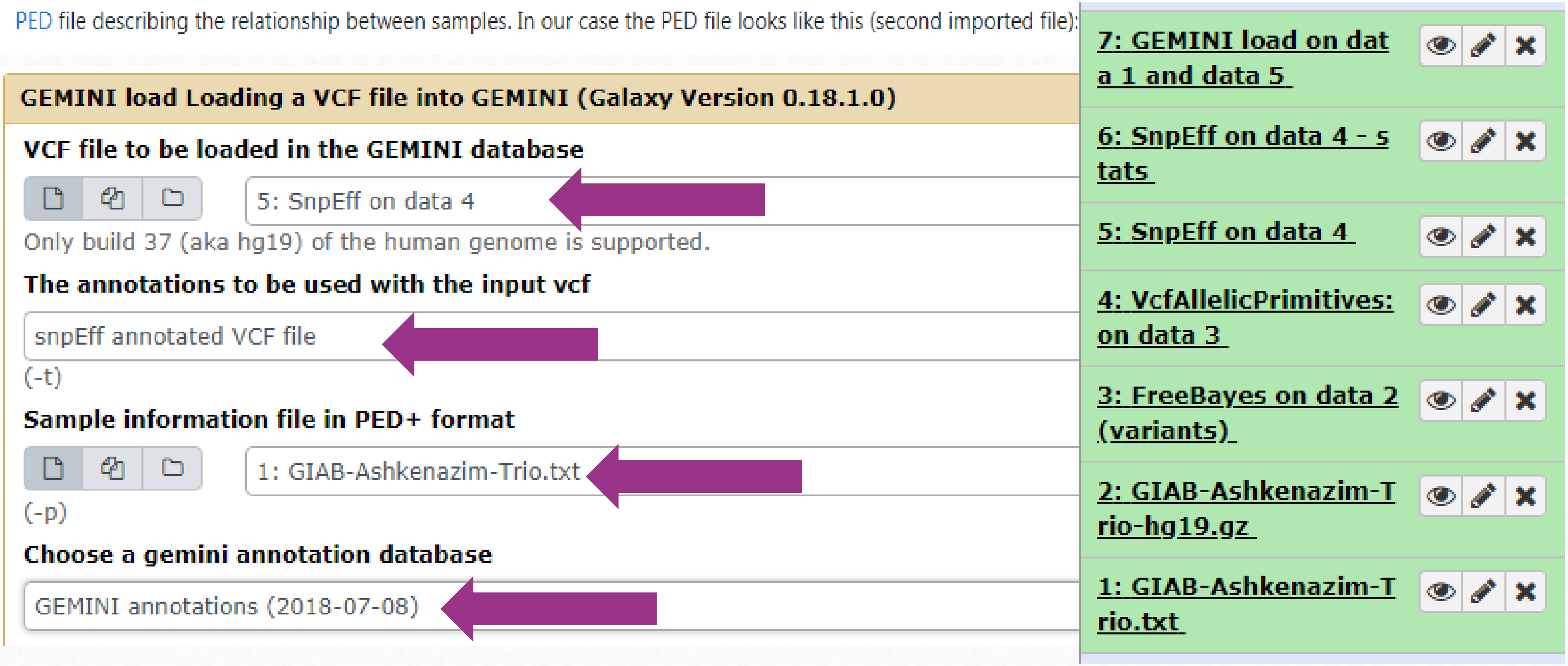
GEMINI annotations (2018-07-08)

1. Hands-On Exome Variant Analysis

Vall d'Hebron Institut de Recerca

Loading data into GEMINI

The first step is to convert a VCF file we would like to analyze into a GEMINI database. For this we will use **GEMINI Load** tool. GEMINI takes as input a VCF file and a PED file describing the relationship between samples. In our case the PED file looks like this (second imported file):



GEMINI load Loading a VCF file into GEMINI (Galaxy Version 0.18.1.0)

VCF file to be loaded in the GEMINI database

5: SnpEff on data 4

Only build 37 (aka hg19) of the human genome is supported.

The annotations to be used with the input vcf

snpEff annotated VCF file

(-t)

Sample information file in PED+ format

1: GIAB-Ashkenazim-Trio.txt

(-p)

Choose a gemini annotation database

GEMINI annotations (2018-07-08)

7: GEMINI load on data 1 and data 5

6: SnpEff on data 4 - stats

5: SnpEff on data 4

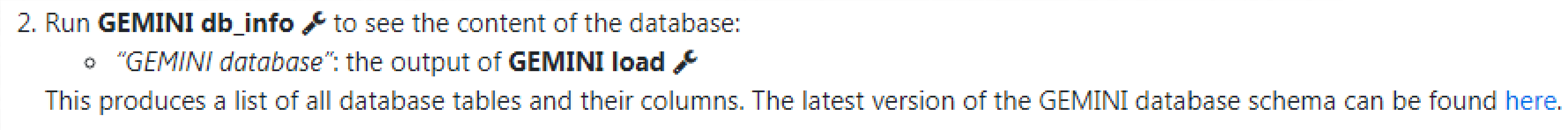
4: VcfAllelicPrimitives: on data 3

3: FreeBayes on data 2 (variants)

2: GIAB-Ashkenazim-Trio-hg19.gz

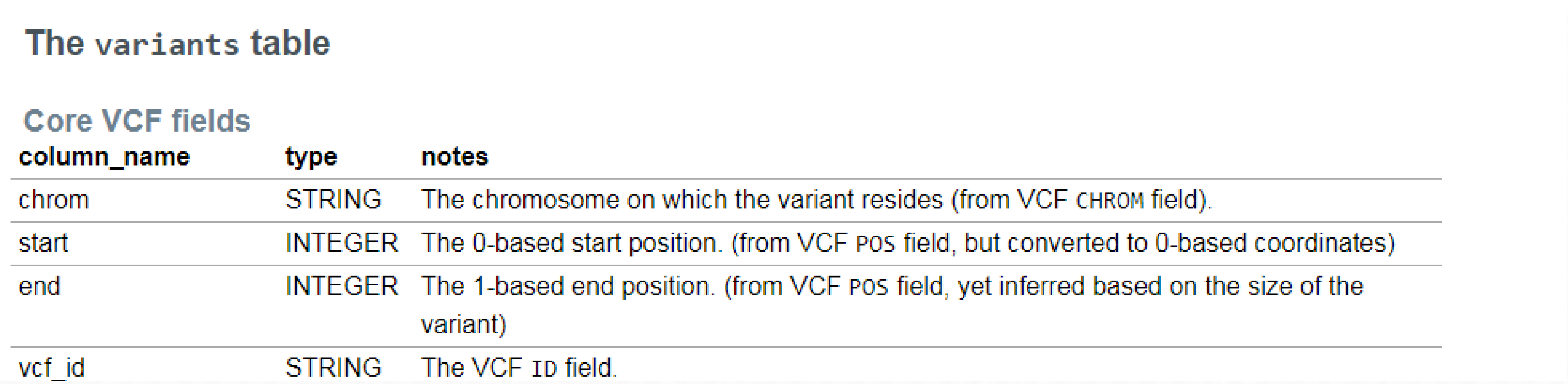
1: GIAB-Ashkenazim-Trio.txt

1. Hands-On Exome Variant Analysis



2. Run **GEMINI db_info** 🔧 to see the content of the database:
    - *"GEMINI database"*: the output of **GEMINI load** 🔧

   This produces a list of all database tables and their columns. The latest version of the GEMINI database schema can be found here.



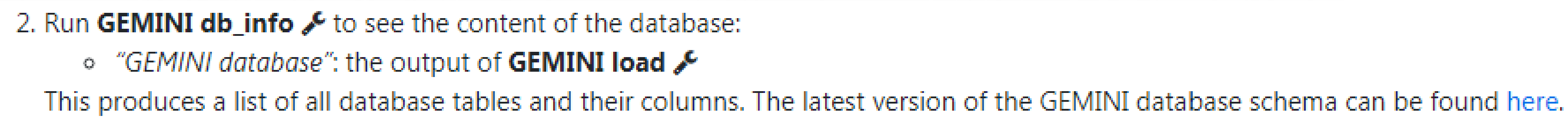
The `variants` table

Core VCF fields

column_name	type	notes
chrom	STRING	The chromosome on which the variant resides (from VCF `CHROM` field).
start	INTEGER	The 0-based start position. (from VCF `POS` field, but converted to 0-based coordinates)
end	INTEGER	The 1-based end position. (from VCF `POS` field, yet inferred based on the size of the variant)
vcf_id	STRING	The VCF `ID` field.

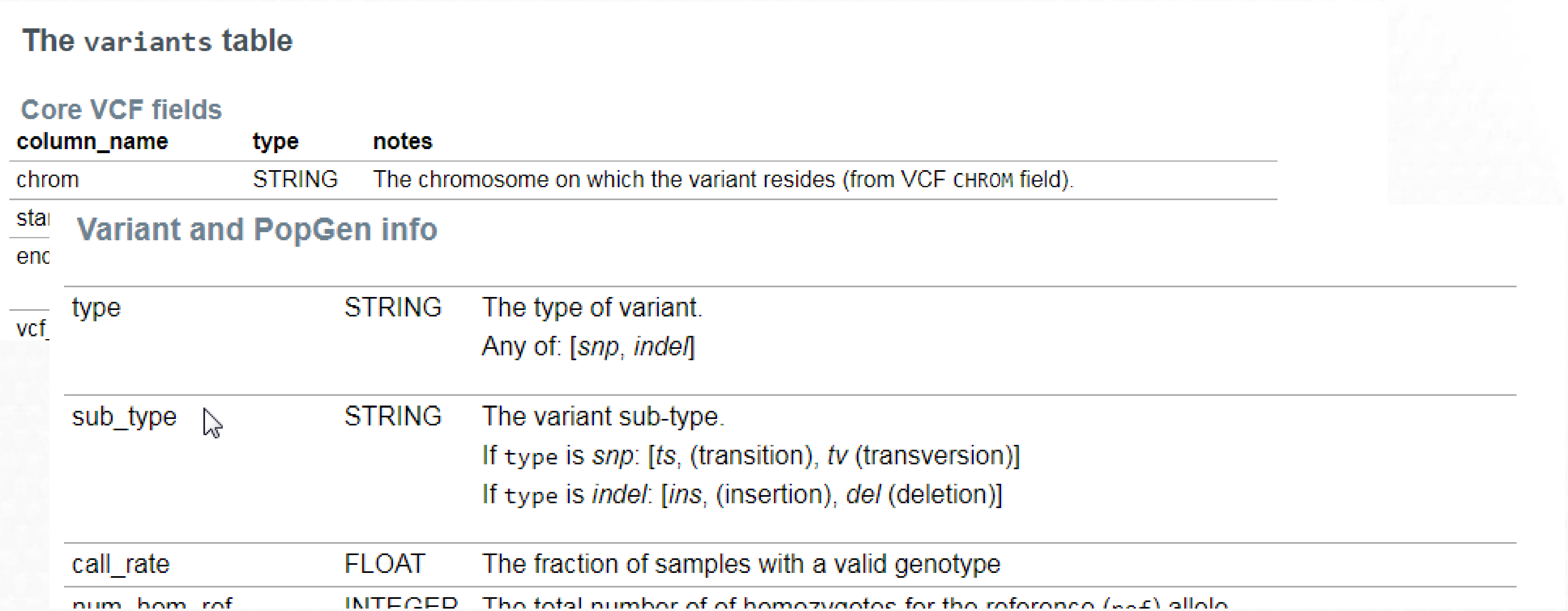
https://gemini.readthedocs.io/en/latest/content/database_schema.html

1. Hands-On Exome Variant Analysis



2. Run **GEMINI db_info** to see the content of the database:
   - *"GEMINI database"*: the output of **GEMINI load**

   This produces a list of all database tables and their columns. The latest version of the GEMINI database schema can be found here.



The `variants` table

Core VCF fields

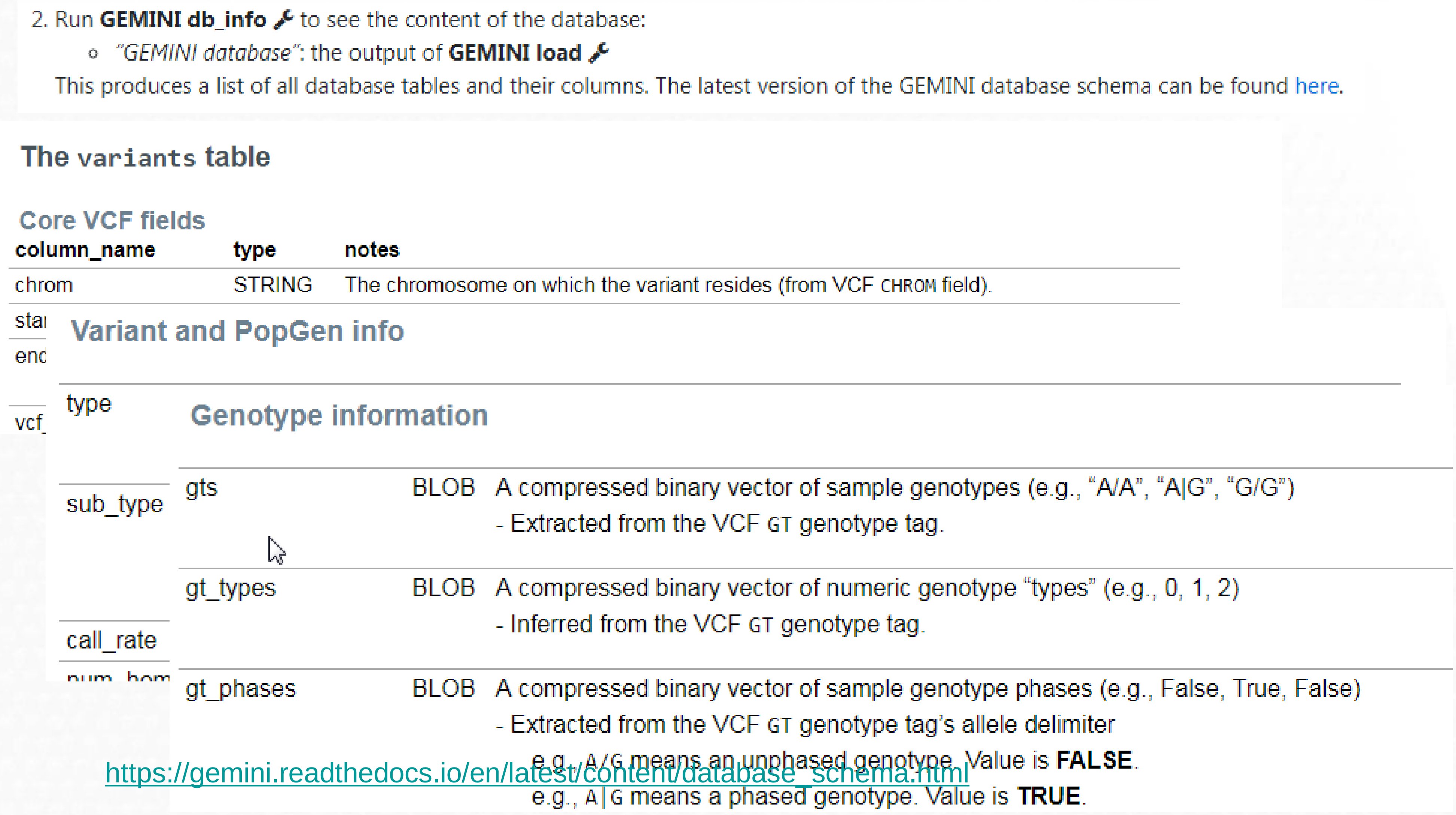
column_name	type	notes
chrom	STRING	The chromosome on which the variant resides (from VCF `CHROM` field).
sta		
enc		
vcf_		

Variant and PopGen info

type	STRING	The type of variant. Any of: [*snp*, *indel*]
sub_type	STRING	The variant sub-type. If `type` is *snp*: [*ts*, (transition), *tv* (transversion)] If `type` is *indel*: [*ins*, (insertion), *del* (deletion)]
call_rate	FLOAT	The fraction of samples with a valid genotype
num_hom_ref	INTEGER	The total number of of homozygotes for the reference (`ref`) allele

https://gemini.readthedocs.io/en/latest/content/database_schema.html

1. Hands-On Exome Variant Analysis



2. Run **GEMINI db_info** 🔧 to see the content of the database:

  ○ *"GEMINI database"*: the output of **GEMINI load** 🔧

This produces a list of all database tables and their columns. The latest version of the GEMINI database schema can be found here.

The `variants` table

Core VCF fields

column_name	type	notes
chrom	STRING	The chromosome on which the variant resides (from VCF `CHROM` field).
sta		
enc		
vcf_		

Variant and PopGen info

type

sub_type

call_rate

num_hom

Genotype information

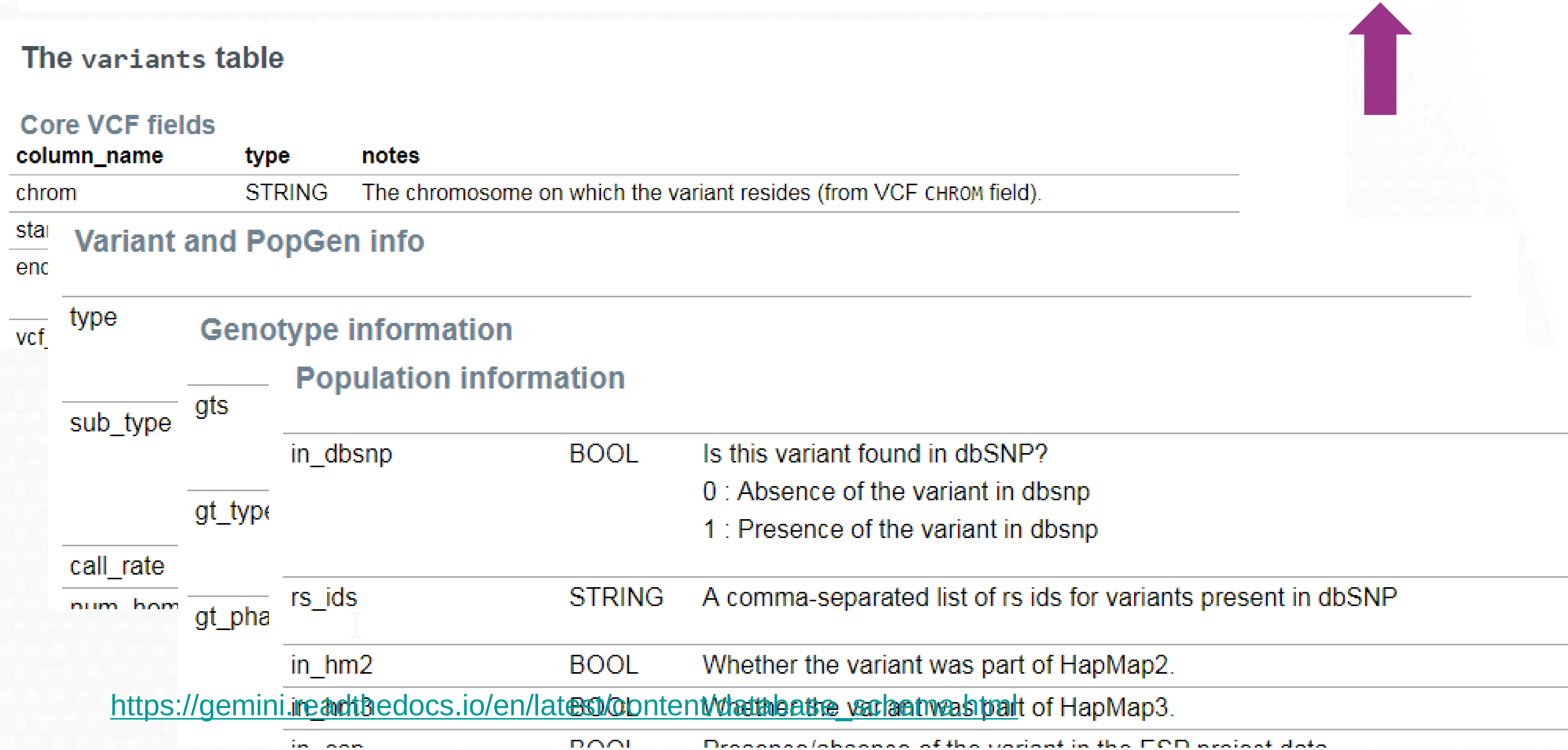
gts	BLOB	A compressed binary vector of sample genotypes (e.g., "A/A", "A|G", "G/G") - Extracted from the VCF `GT` genotype tag.
gt_types	BLOB	A compressed binary vector of numeric genotype "types" (e.g., 0, 1, 2) - Inferred from the VCF `GT` genotype tag.
gt_phases	BLOB	A compressed binary vector of sample genotype phases (e.g., False, True, False) - Extracted from the VCF `GT` genotype tag's allele delimiter e.g., `A/G` means an unphased genotype. Value is **FALSE**. e.g., `A|G` means a phased genotype. Value is **TRUE**.

https://gemini.readthedocs.io/en/latest/content/database_schema.html

1. Hands-On Exome Variant Analysis

2. Run **GEMINI db_info** 🔧 to see the content of the database:
    - *"GEMINI database"*: the output of **GEMINI load** 🔧

    This produces a list of all database tables and their columns. The latest version of the GEMINI database schema can be found here.



The variants table

Core VCF fields

column_name	type	notes
chrom	STRING	The chromosome on which the variant resides (from VCF CHROM field).
stai		
enc		
vcf_		

Variant and PopGen info

type		
sub_type		
call_rate		
num_hom		

Genotype information

gts		
gt_type		
gt_pha		

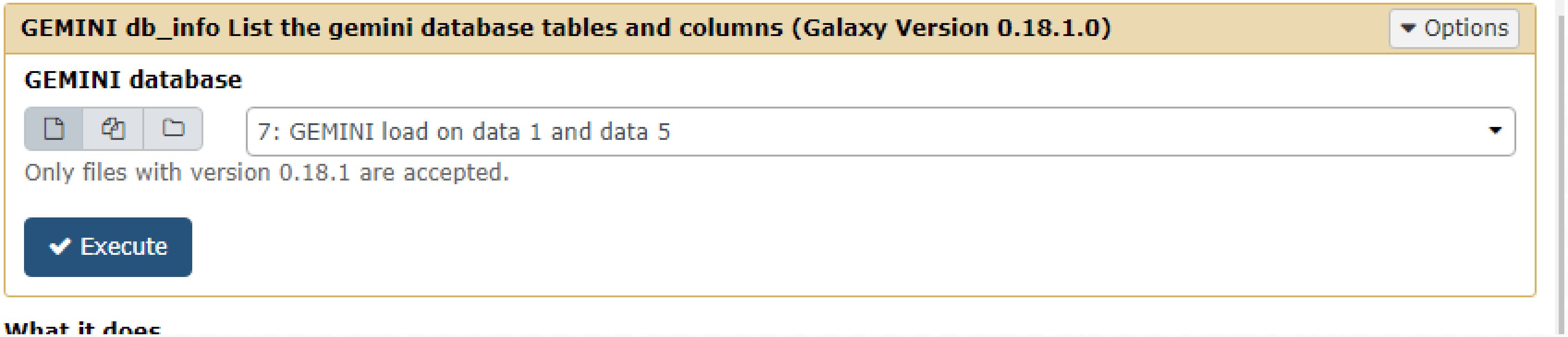
Population information

in_dbsnp	BOOL	Is this variant found in dbSNP? 0 : Absence of the variant in dbsnp 1 : Presence of the variant in dbsnp
rs_ids	STRING	A comma-separated list of rs ids for variants present in dbSNP
in_hm2	BOOL	Whether the variant was part of HapMap2.
in_hm3	BOOL	Whether the variant was part of HapMap3.
in_esp	BOOL	Presence/absence of the variant in the ESP project data

https://gemini.readthedocs.io/en/latest/content/database_schema.html

1. Hands-On Exome Variant Analysis

GEMINI db_info List the gemini database tables and columns



GEMINI db_info List the gemini database tables and columns (Galaxy Version 0.18.1.0)

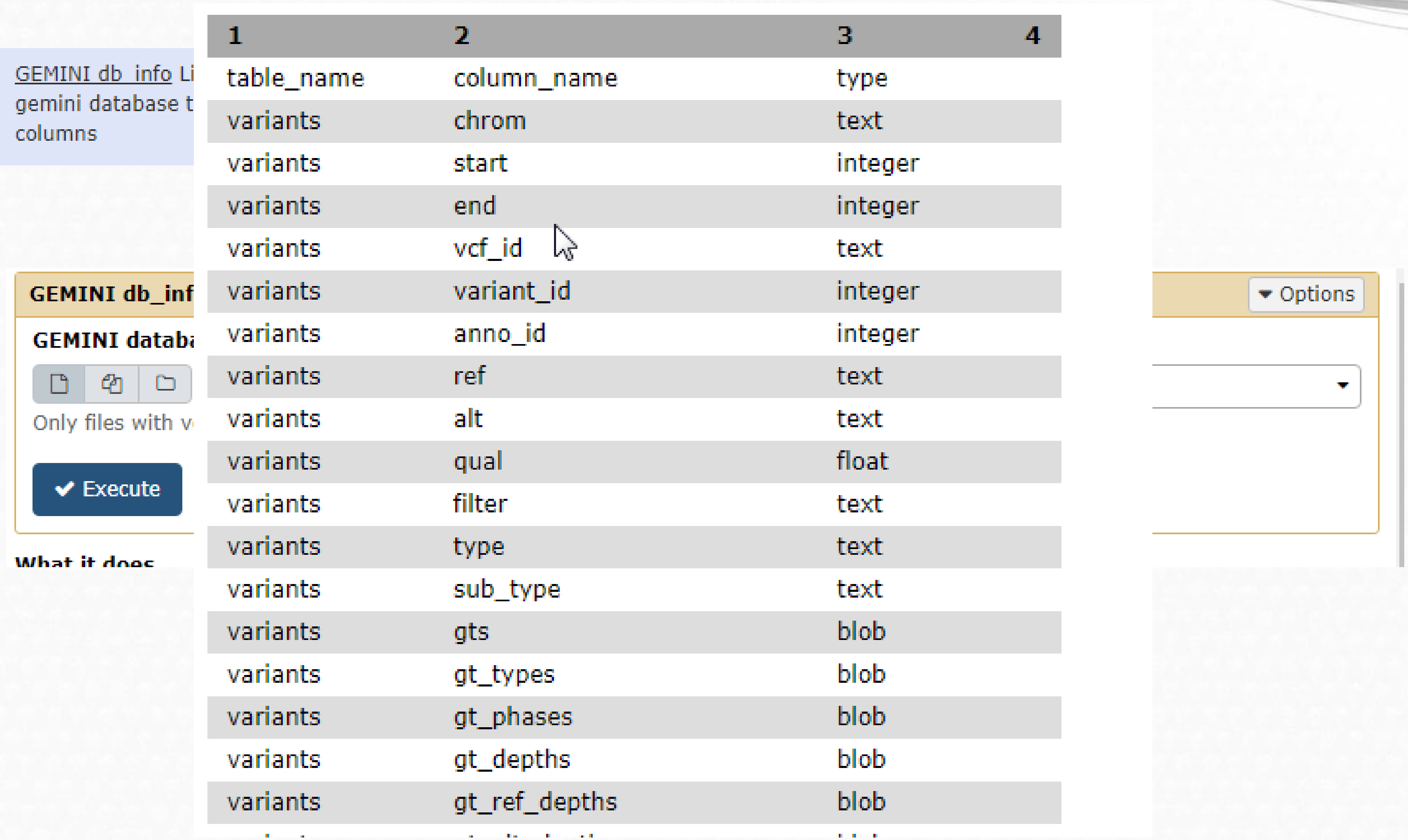
GEMINI database

7: GEMINI load on data 1 and data 5

Only files with version 0.18.1 are accepted.

What it does

1. Hands-On Exome Variant Analysis



GEMINI db_info Li
gemini database t
columns

GEMINI db_inf

GEMINI databa

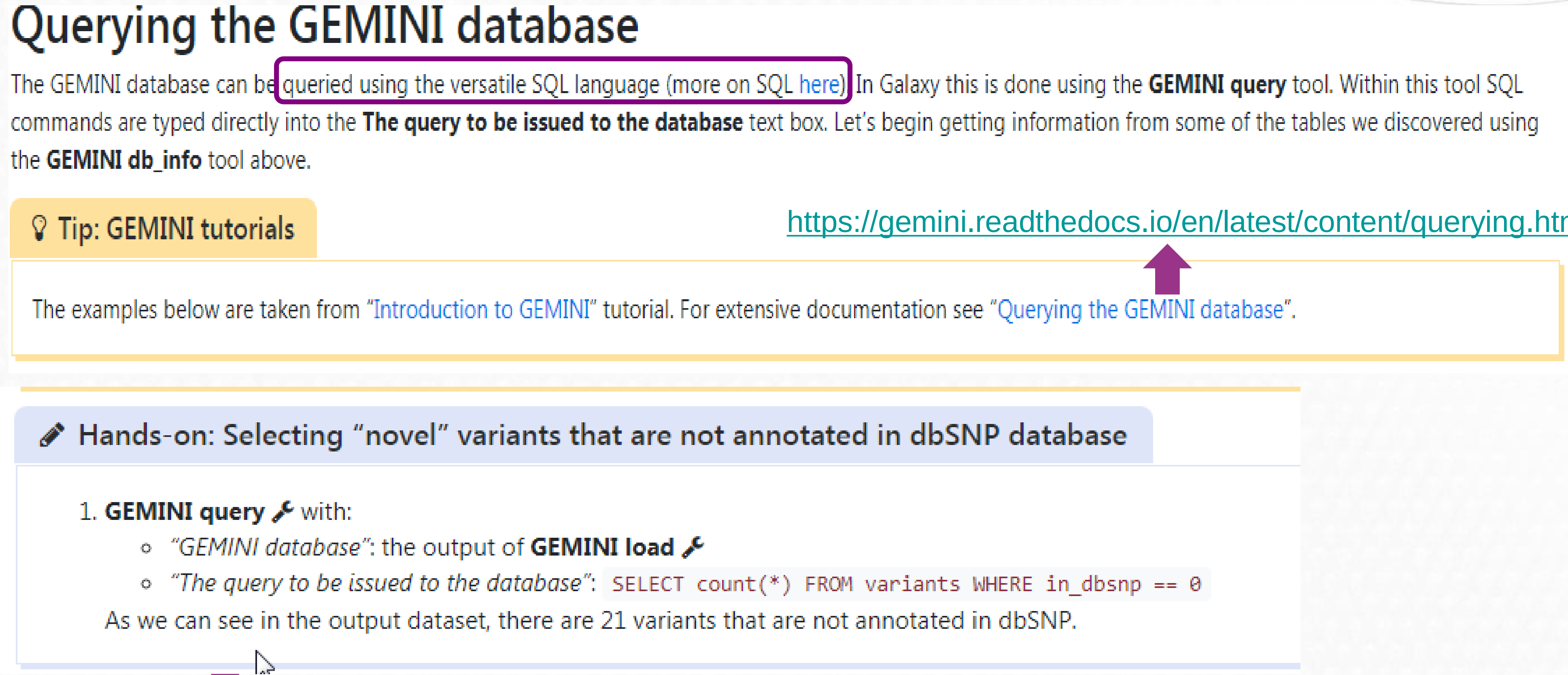
Only files with v

✔ Execute

Options

What it does

1	2	3	4
table_name	column_name	type	
variants	chrom	text	
variants	start	integer	
variants	end	integer	
variants	vcf_id	text	
variants	variant_id	integer	
variants	anno_id	integer	
variants	ref	text	
variants	alt	text	
variants	qual	float	
variants	filter	text	
variants	type	text	
variants	sub_type	text	
variants	gts	blob	
variants	gt_types	blob	
variants	gt_phases	blob	
variants	gt_depths	blob	
variants	gt_ref_depths	blob	



Querying the GEMINI database

The GEMINI database can be queried using the versatile SQL language (more on SQL here). In Galaxy this is done using the **GEMINI query** tool. Within this tool SQL commands are typed directly into the **The query to be issued to the database** text box. Let's begin getting information from some of the tables we discovered using the **GEMINI db_info** tool above.

https://gemini.readthedocs.io/en/latest/content/querying.html

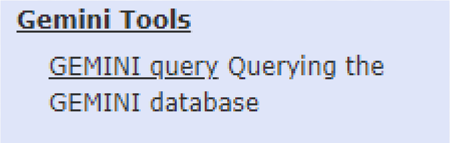
Tip: GEMINI tutorials

The examples below are taken from "Introduction to GEMINI" tutorial. For extensive documentation see "Querying the GEMINI database".

Hands-on: Selecting "novel" variants that are not annotated in dbSNP database

1. **GEMINI query** with:
   - *"GEMINI database"*: the output of **GEMINI load**
   - *"The query to be issued to the database"*: `SELECT count(*) FROM variants WHERE in_dbsnp == 0`

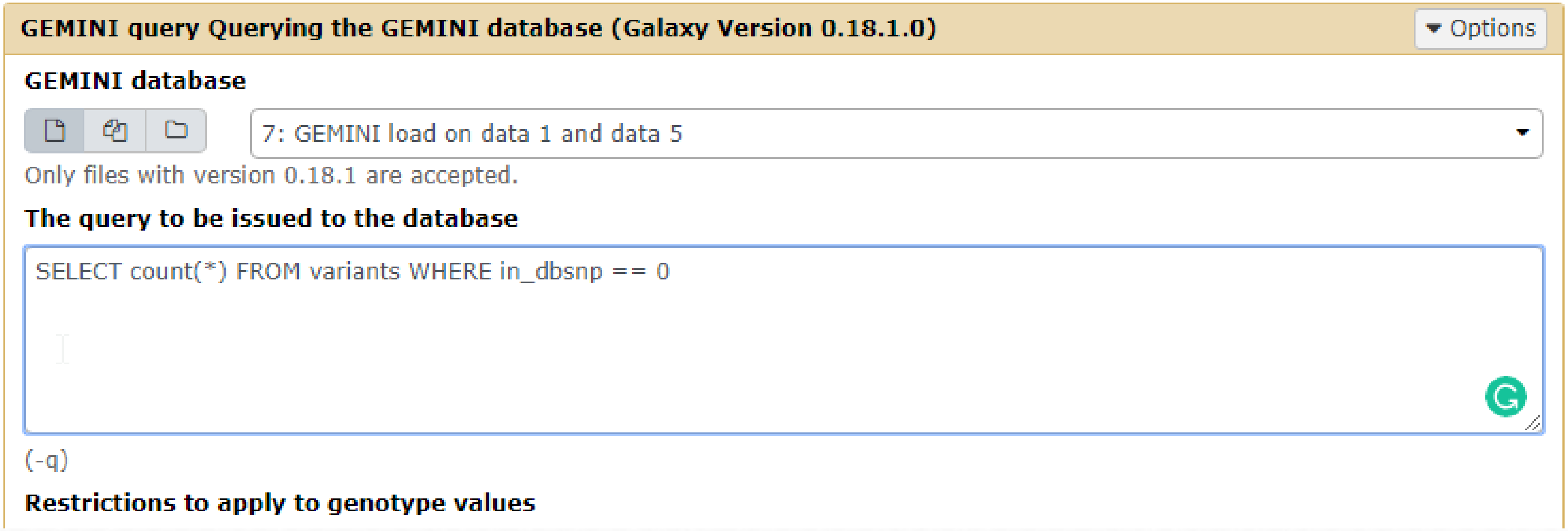
   As we can see in the output dataset, there are 21 variants that are not annotated in dbSNP.



Gemini Tools

GEMINI query Querying the GEMINI database

1. Hands-On Exome Variant Analysis



GEMINI query Querying the GEMINI database (Galaxy Version 0.18.1.0)

GEMINI database

7: GEMINI load on data 1 and data 5

Only files with version 0.18.1 are accepted.

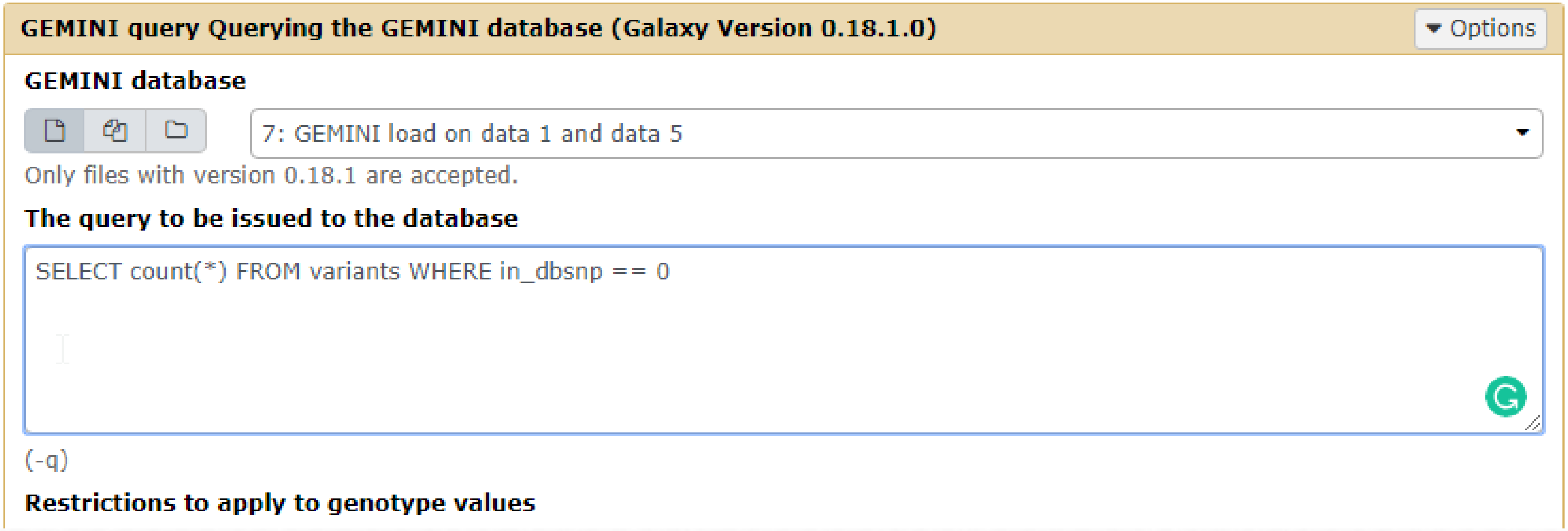
The query to be issued to the database

```
SELECT count(*) FROM variants WHERE in_dbsnp == 0
```

(-q)

Restrictions to apply to genotype values

1. Hands-On Exome Variant Analysis



GEMINI query Querying the GEMINI database (Galaxy Version 0.18.1.0)

GEMINI database

7: GEMINI load on data 1 and data 5

Only files with version 0.18.1 are accepted.

The query to be issued to the database

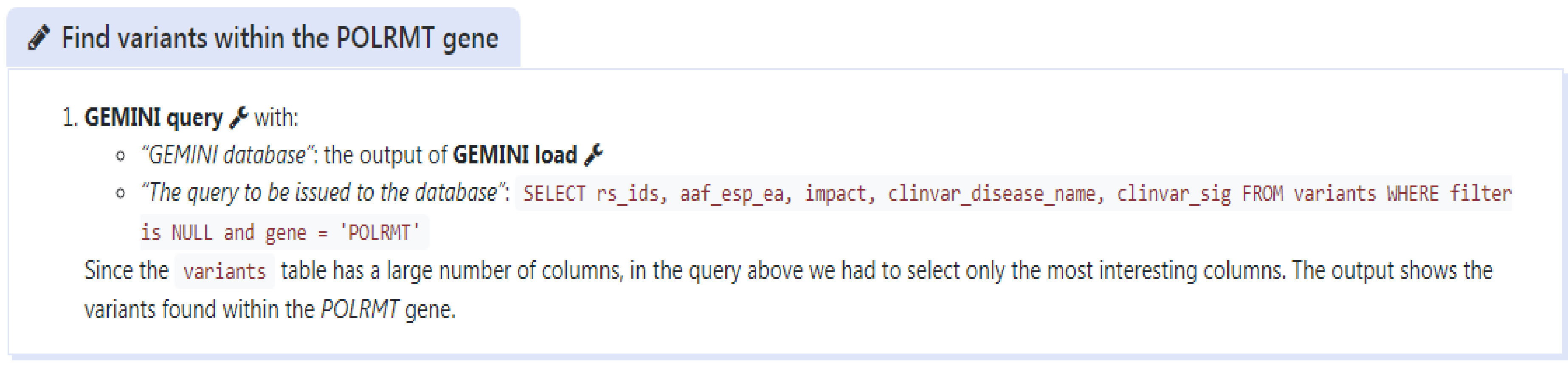
```
SELECT count(*) FROM variants WHERE in_dbsnp == 0
```

(-q)

Restrictions to apply to genotype values

1
21

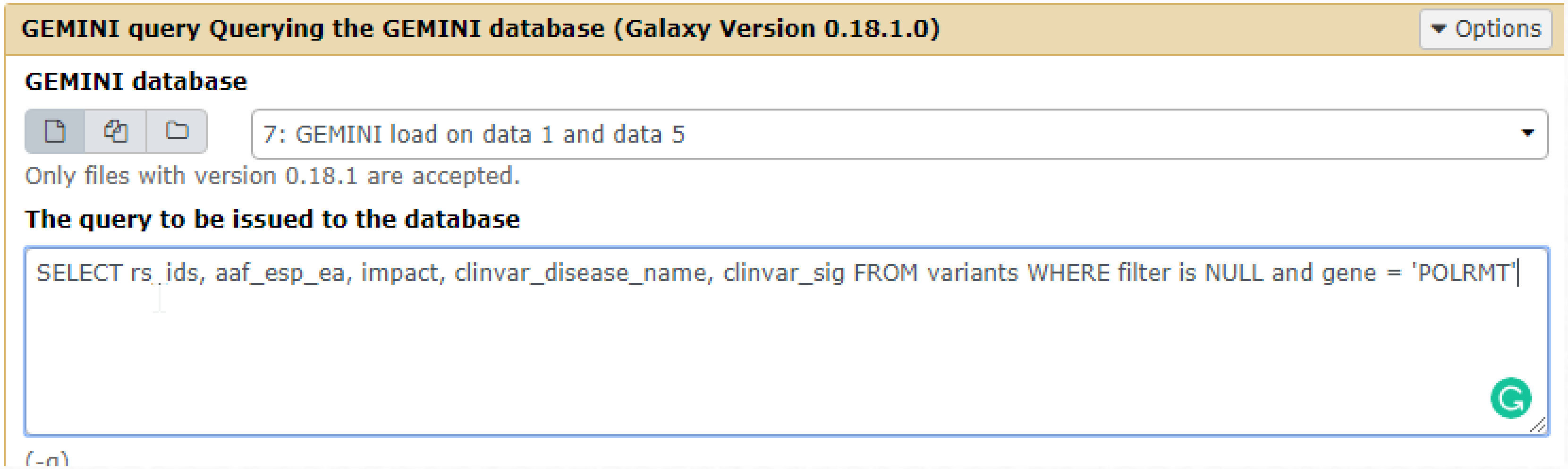
1. Hands-On Exome Variant Analysis



✎ Find variants within the POLRMT gene

1. **GEMINI query** 🔧 with:
   - *"GEMINI database"*: the output of **GEMINI load** 🔧
   - *"The query to be issued to the database"*: `SELECT rs_ids, aaf_esp_ea, impact, clinvar_disease_name, clinvar_sig FROM variants WHERE filter is NULL and gene = 'POLRMT'`

   Since the `variants` table has a large number of columns, in the query above we had to select only the most interesting columns. The output shows the variants found within the *POLRMT* gene.



GEMINI query Querying the GEMINI database (Galaxy Version 0.18.1.0)

GEMINI database

7: GEMINI load on data 1 and data 5

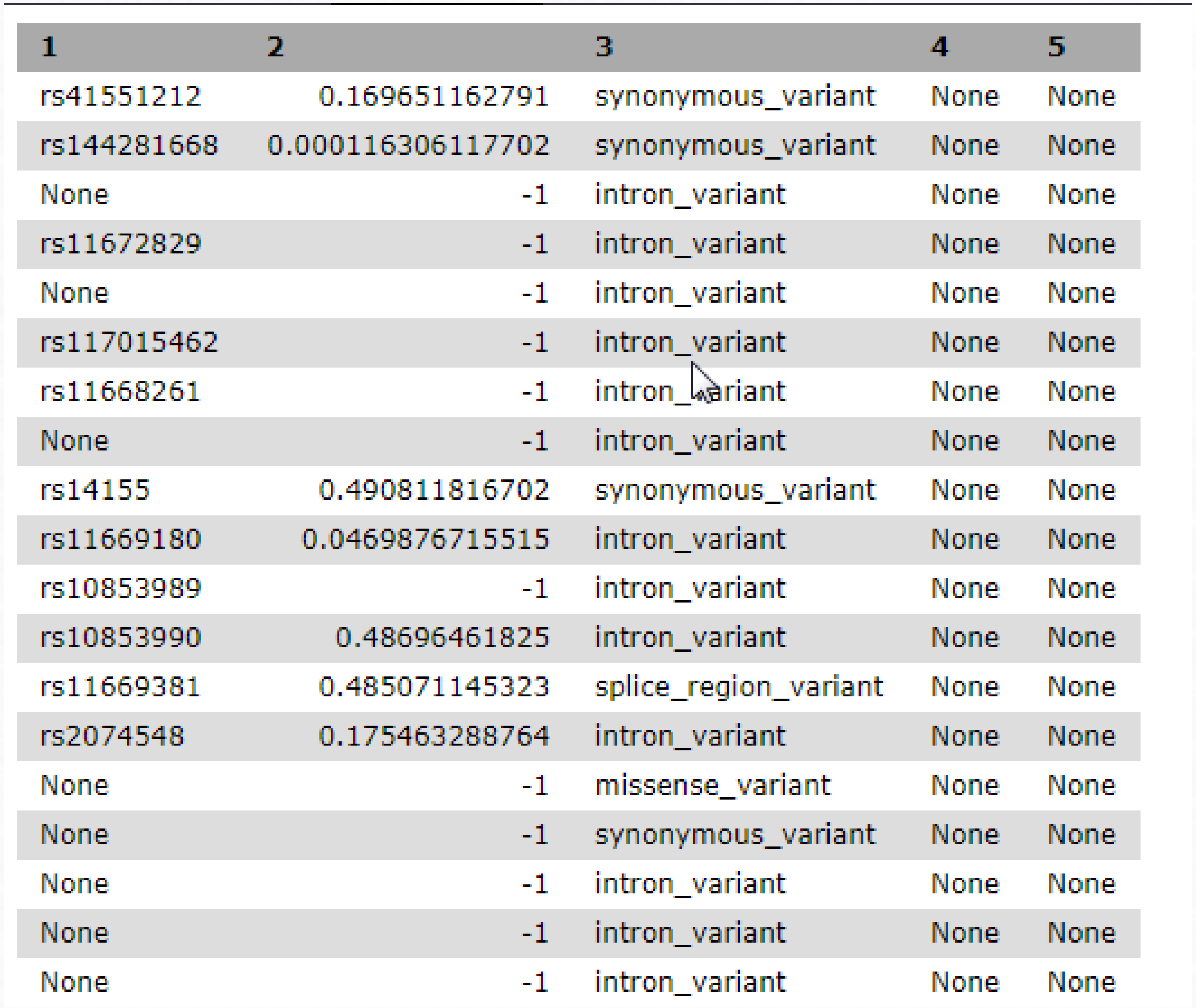
Only files with version 0.18.1 are accepted.

The query to be issued to the database

```
SELECT rs_ids, aaf_esp_ea, impact, clinvar_disease_name, clinvar_sig FROM variants WHERE filter is NULL and gene = 'POLRMT'
```

(-q)

1. Hands-On Exome Variant Analysis



1	2	3	4	5
rs41551212	0.169651162791	synonymous_variant	None	None
rs144281668	0.000116306117702	synonymous_variant	None	None
None	-1	intron_variant	None	None
rs11672829	-1	intron_variant	None	None
None	-1	intron_variant	None	None
rs117015462	-1	intron_variant	None	None
rs11668261	-1	intron_variant	None	None
None	-1	intron_variant	None	None
rs14155	0.490811816702	synonymous_variant	None	None
rs11669180	0.0469876715515	intron_variant	None	None
rs10853989	-1	intron_variant	None	None
rs10853990	0.48696461825	intron_variant	None	None
rs11669381	0.485071145323	splice_region_variant	None	None
rs2074548	0.175463288764	intron_variant	None	None
None	-1	missense_variant	None	None
None	-1	synonymous_variant	None	None
None	-1	intron_variant	None	None
None	-1	intron_variant	None	None
None	-1	intron_variant	None	None