# Identification of somatic and germline variants from tumor and normal sample pairs

Somatic variants tutorial

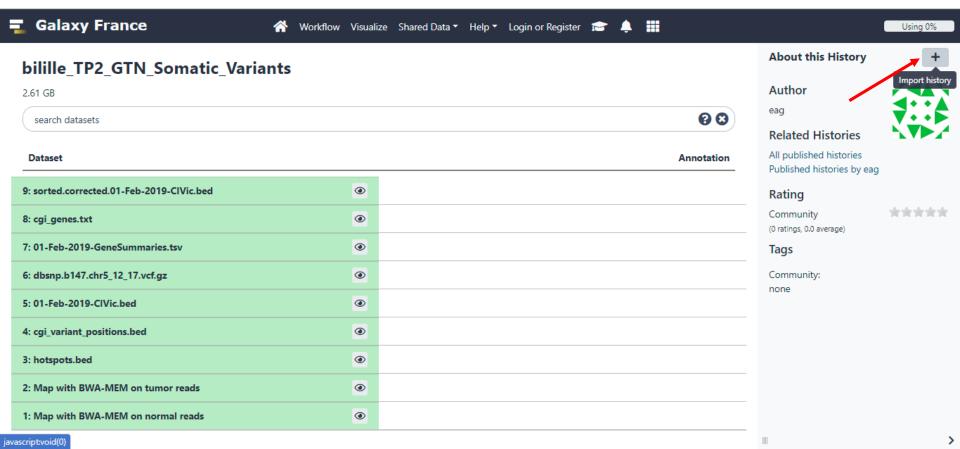


# Workflow

- 1. Mapped reads postprocessing
  - a. Filtering on mapped reads properties
  - b. Removing duplicate reads
  - c. Left-align reads around indels
  - d. Recalibrate read mapping qualities
  - e. Refilter reads based on mapping quality
- 2. Variant calling and classification
- 3. Variant annotation and reporting
  - a. Adding annotations to the called variants
  - b. Reporting selected subsets of variants
  - c. Generating reports of genes affected by variants
  - d. Adding additional annotations to the gene-centered report

# <u>Starting from BAMs : Import Shared History</u>

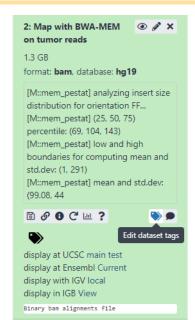
## Shared Data → Histories

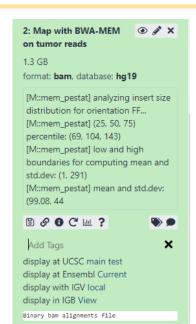


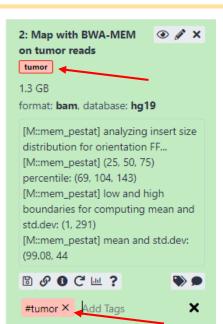
# ↑ Tip: Adding a tag □

## Prepare Data

- Click on the dataset
- Click on Edit dataset tags
- Add a tag starting with #
   Tags starting with # will be automatically propagated to the outputs of tools using this dataset.
- · Check that the tag is appearing below the dataset name





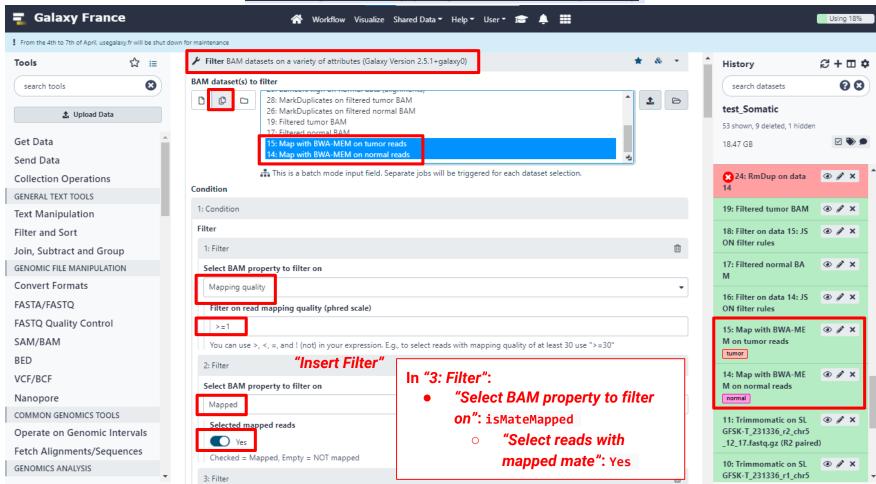


1. Mapped reads postprocessing

# 1. Mapped reads postprocessing

Filtering on mapped reads properties

# Filtering for mapping status and quality



# Filtering for mapping status and quality

There is not only one tool that can filter reads.

*To Do*: find another tool in Galaxy to perform the same operation

# Filtering for mapping status and quality

There is not only one tool that can filter reads.

To Do: find another tool in Galaxy to perform the same operation

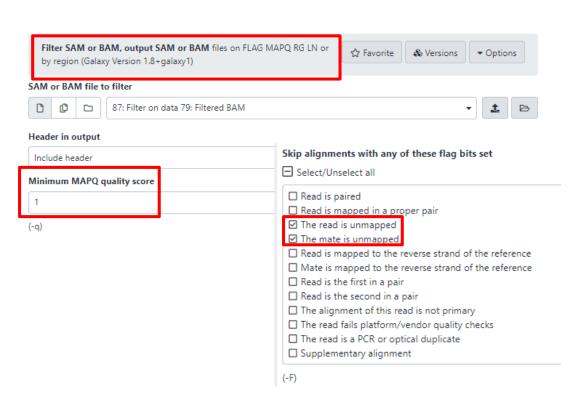


Filter SAM or BAM, output SAM or BAM based on samtools view

equivalent to

Filter BAM datasets on a variety of attributes

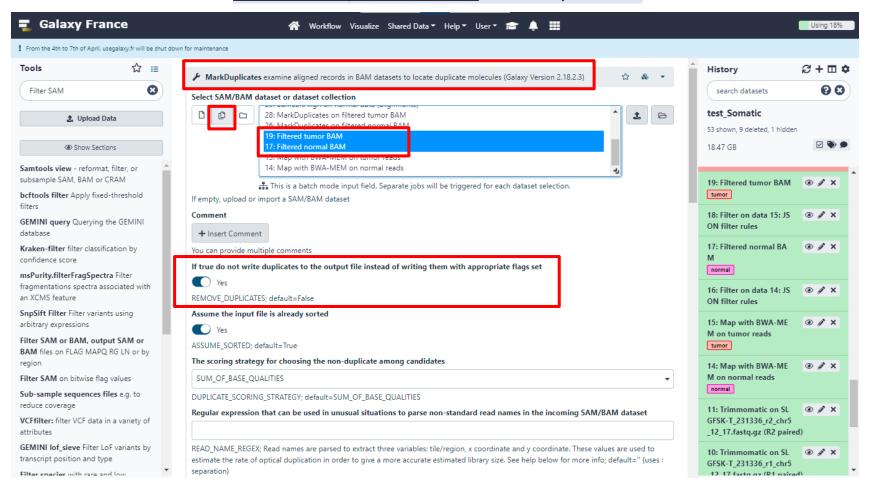
Based on bamtools filter



Mapped reads postprocessing

Removing duplicate reads

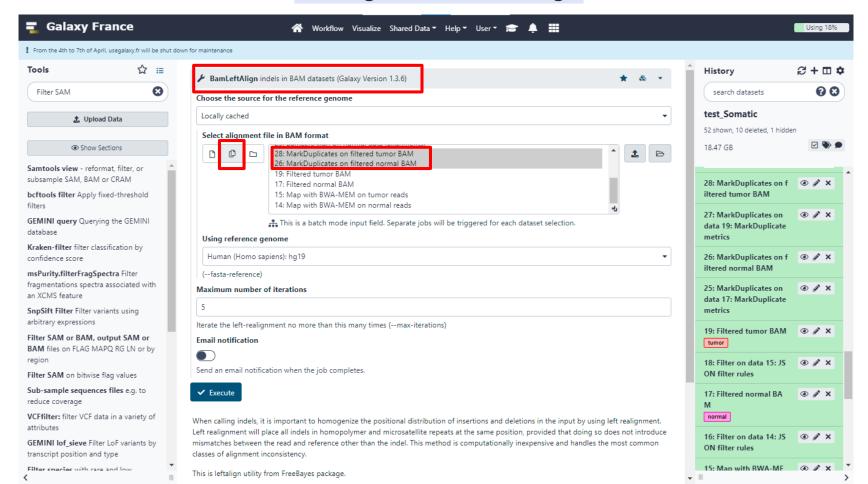
# Remove duplicates with MarkDuplicates



# Mapped reads postprocessing

Left-align reads around indels

# Left-align with BamLeftAlign

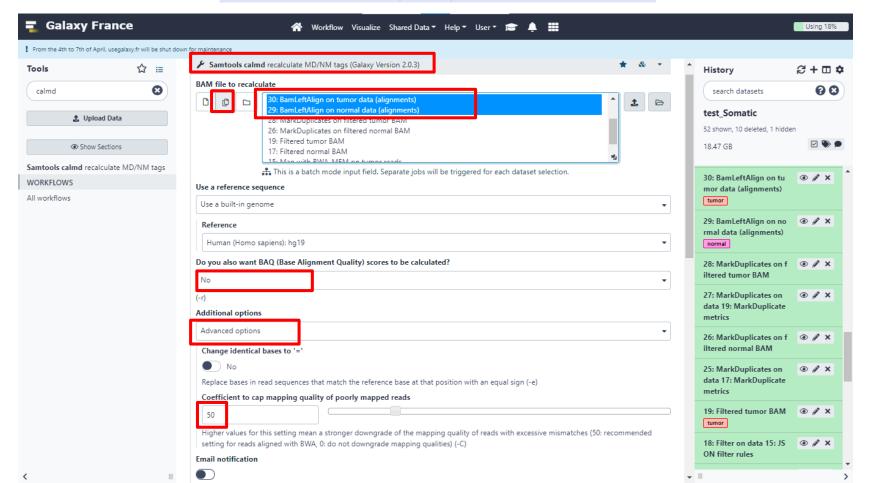


# WHY LEFT ALIGN???

# Mapped reads postprocessing

Recalibrate read mapping qualities

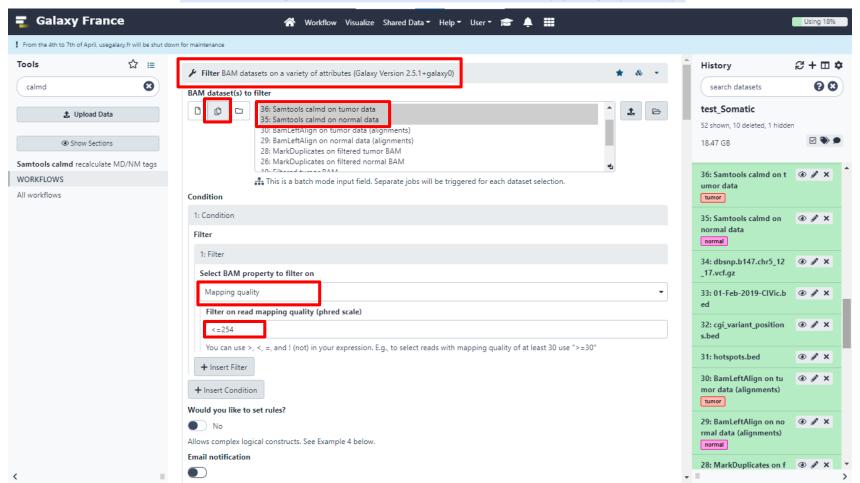
# Recalibrate read quality scores with CalMD



# Mapped reads postprocessing

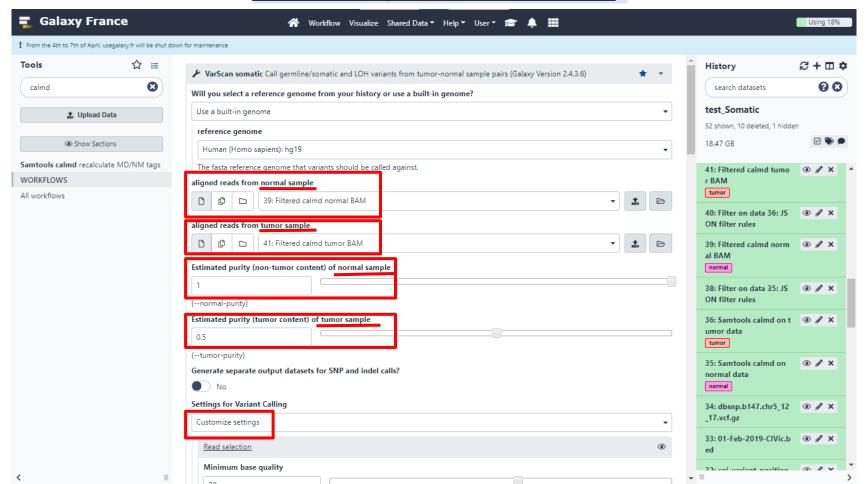
e. Refilter reads based on mapping quality

# Eliminating reads with undefined mapping quality

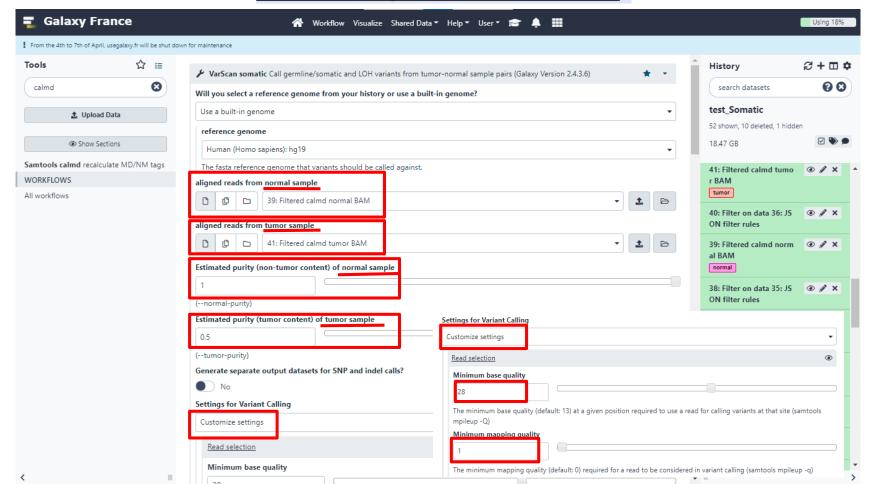


2. Variant calling and classification

# Variant calling with VarScan somatic



# Variant calling with VarScan somatic



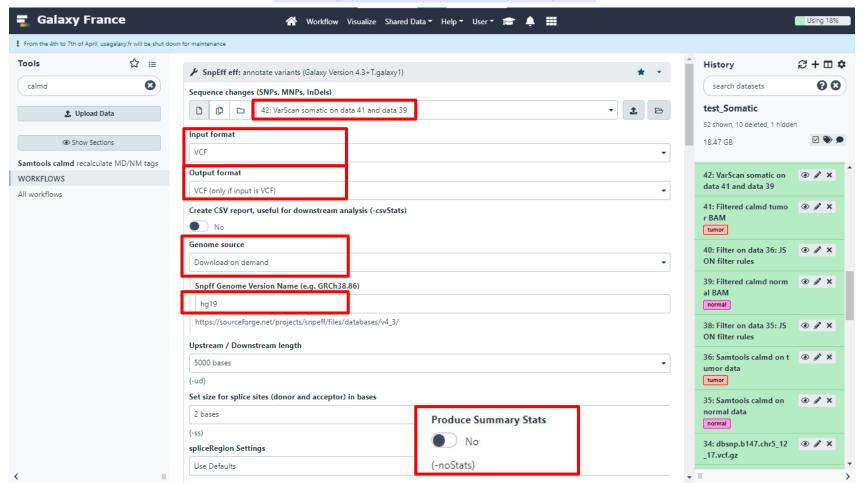
Variant annotation and reporting

# Adding annotations to the called variants

a. Adding annotations to the called variants

Adding functional genomic annotations

# Adding annotations with SnpEff

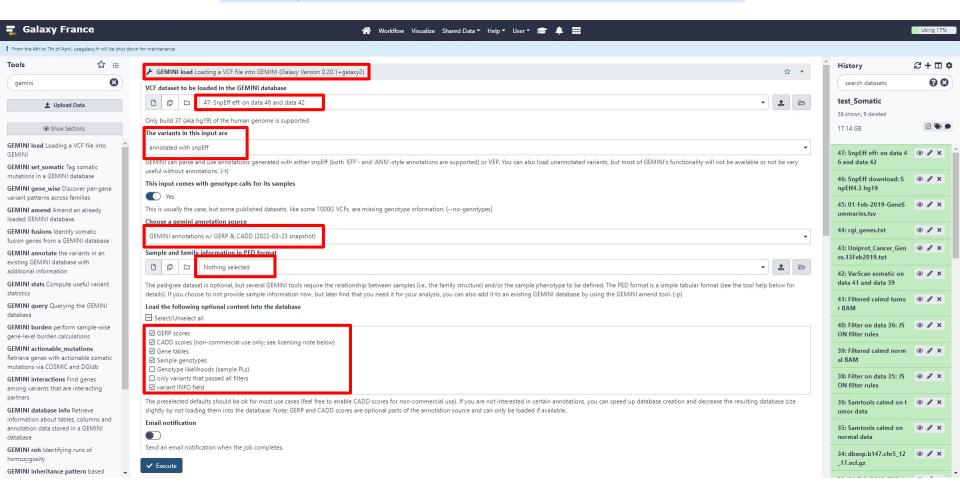


# Adding genetic and clinical evidence-based

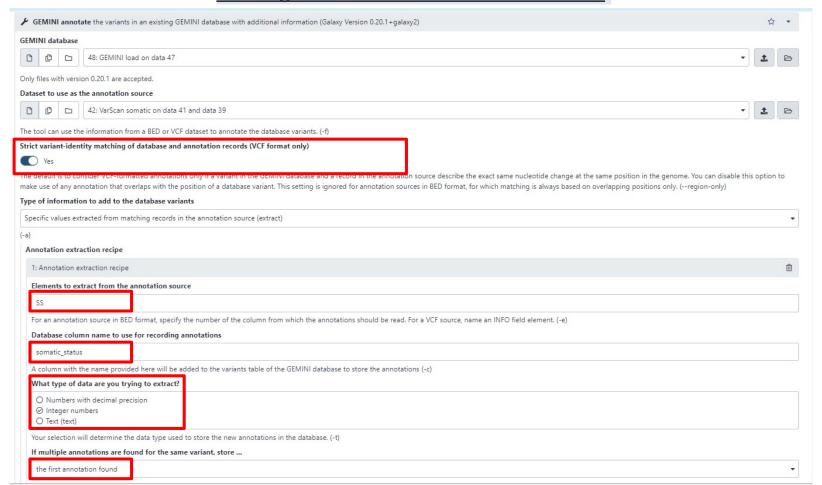
Adding annotations to the called variants

annotations

# <u>Creating a GEMINI database from a variants dataset</u>



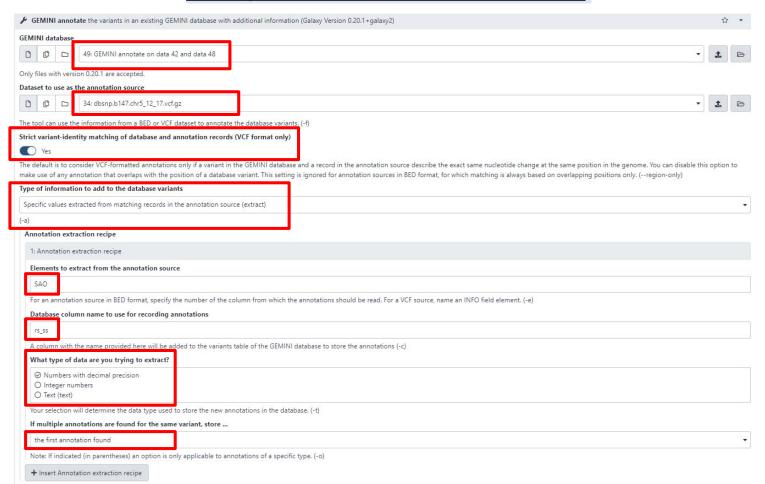
# Making variant call statistics accessible



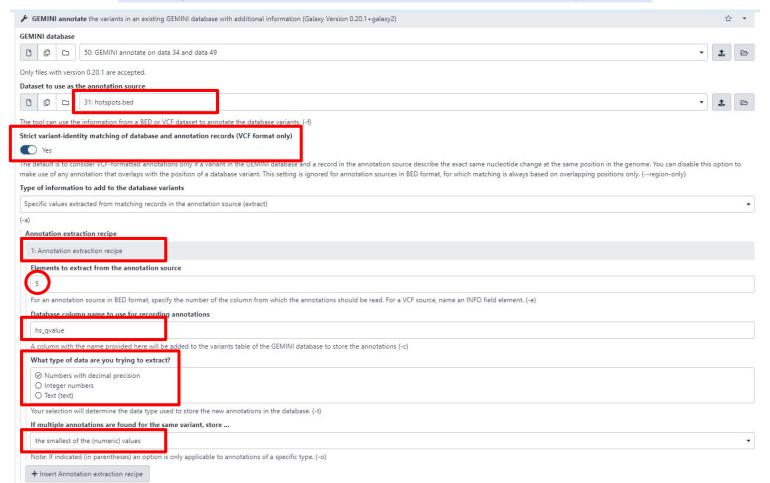
# Making variant call statistics accessible

	2: Annotation extraction recipe	Û
ſ	Elements to extract from the annotation source	
ı	GPV	
	For an annotation source in BED format, specify the number of the column from which the annotations should be read. For a VCF source, name an INFO field element. (-e)	
	Database column name to use for recording annotations	
	germline_p	
	A column with the name provided here will be added to the variants table of the GEMINI database to store the annotations (-c)	
ſ	What type of data are you trying to extract?	
	Numbers with decimal precision Integer numbers Text (text)	
-	Your selection will determine the data type used to store the new annotations in the database. (-t)	
	If multiple annotations are found for the same variant, store	
	the first annotation found	•
•	Note: If indicated (in parentheses) an option is only applicable to annotations of a specific type. (-o)	
	3: Annotation extraction recipe	Û
ſ	Elements to extract from the annotation source	
ı	SPV	
	For an annotation source in BED format, specify the number of the column from which the annotations should be read. For a VCF source, name an INFO field element. (-e)	
	Database column name to use for recording annotations	
	somatic_p	
	A column with the name provided here will be added to the variants table of the GEMINI database to store the annotations (-c)	
	What type of data are you trying to extract?	
	○ Integer numbers ○ Text (text)	
	Your selection will determine the data type used to store the new annotations in the database. (-t)	
	If multiple annotations are found for the same variant, store	
	the first annotation found	

# Adding further annotations from dbSNP



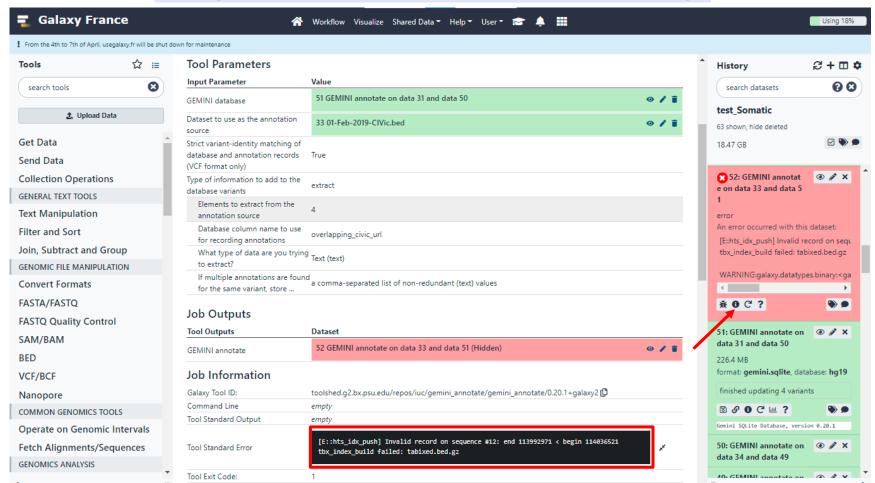
# Adding further annotations from Cancer Hotspots v2



# Adding links to CIViC

GEMINI annotate the variants in an existing GEMINI database with additional information (Galaxy Version 0.20.1+galaxy2)	☆	*
GEMINI database		
□   □     □   □     51: GEMINI annotate on data 31 and data 50	1	
Only files with version 0.20.1 are accepted.		
Dataset to use as the annotation source		
□         □         33: 01-Feb-2019-CIVic.bed	±	
The tool can use the information from a BED or VCF dataset to annotate the database variants. (-f)		
Strict variant-identity matching of database and annotation records (VCF format only)   Yes		
The default is to consider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annotation source describe the exact same nucleotide change at the same position in the genome. You can disable this make use of any annotation that overlaps with the position of a database variant. This setting is ignored for annotation sources in BED format, for which matching is always based on overlapping positions only. (region-only)	option	to
Type of information to add to the database variants		
Specific values extracted from matching records in the annotation source (extract)		-
(-a)		
Annotation extraction recipe		
1: Annotation extraction recipe		
Elements to extract from the annotation source		
4		
For an annotation source in BED format, specify the number of the column from which the annotations should be read. For a VCF source, name an INFO field element. (-e)		
Database column name to use for recording annotations		
overlapping_civic_url		
A column with the name provided here will be added to the variants table of the GEMINI database to store the annotations (-c)		
What type of data are you trying to extract?		
O Numbers with decimal precision O Integer numbers ⊙ Text (text)		
Your selection will determine the data type used to store the new annotations in the database. (-t)		
If multiple annotations are found for the same variant, store		
a comma-separated list of non-redundant (text) values		•
Note: If indicated (in parentheses) an option is only applicable to annotations of a specific type. (-o)		
+ Insert Annotation extraction recipe		

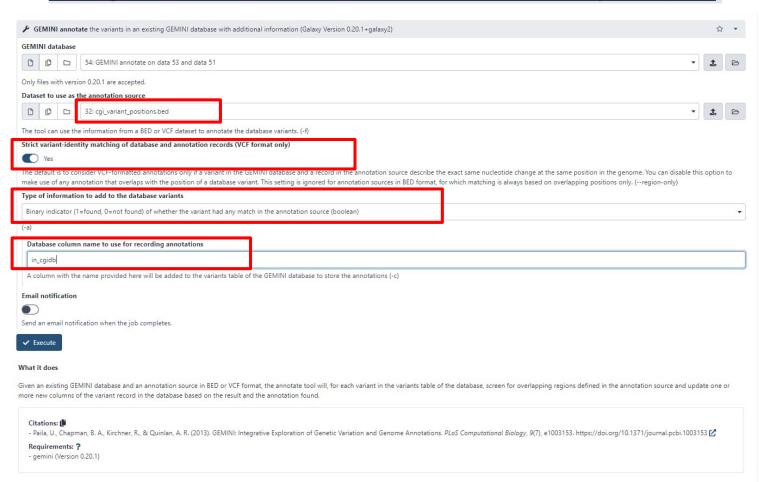
# Adding links to CIViC: How to know what went wrong?



# Adding links to CIViC

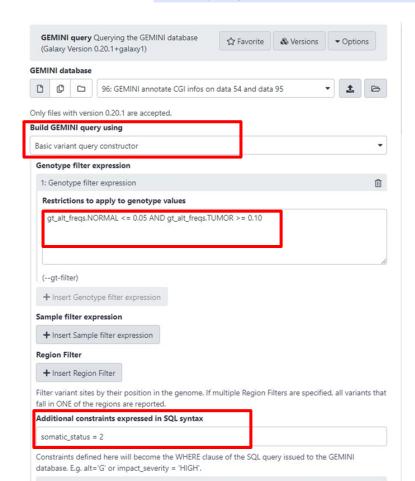
F GEMINI annotate the variants in an existing GEMINI database with additional information (Galaxy Version 0.20.1+galaxy2)		☆	-		
GEMINI database					
51: GEMINI annotate on data 31 and data 50	,	1			
Only files with version 0.20.1 are accepted.					
Dataset to use as the annotation source					
53: sorted.corrected.01-Feb-2019-CIVic.bed	, 1	L			
The tool can use the information from a BED or VCF dataset to annotate the database variants. (-f)					
Strict variant-identity matching of database and annotation records (VCF format only)					
The default is to consider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annotation source describe the exact same nucleotide change at the same position in the genome. You can disable make use of any annotation that overlaps with the position of a database variant. This setting is ignored for annotation sources in BED format, for which matching is always based on overlapping positions only. (region-only)	:his op	tion t	:0		
Type of information to add to the database variants					
Specific values extracted from matching records in the annotation source (extract)			*		
(-a)					
Annotation extraction recipe					
1: Annotation extraction recipe					
Elements to extract from the annotation source					
4					
For an annotation source in BED format, specify the number of the column from which the annotations should be read. For a VCF source, name an INFO field element. (-e)					
Database column name to use for recording annotations					
overlapping_civic_url					
A column with the name provided here will be added to the variants table of the GEMINI database to store the annotations (-c)					
What type of data are you trying to extract?					
O Numbers with decimal precision O Integer numbers  ⊘ Text (text)					
Your selection will determine the data type used to store the new annotations in the database. (-t)					
If multiple annotations are found for the same variant, store					
a comma-separated list of non-redundant (text) values					
Natural findicated (in pagestheses) an entire is only applicable to appetations of a procific type (in)					

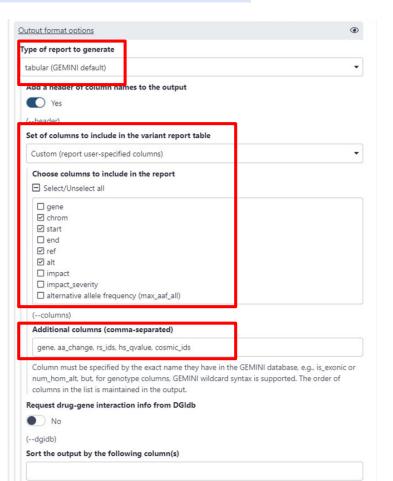
# Adding further annotations from Cancer Genome Interpreter (CGI)



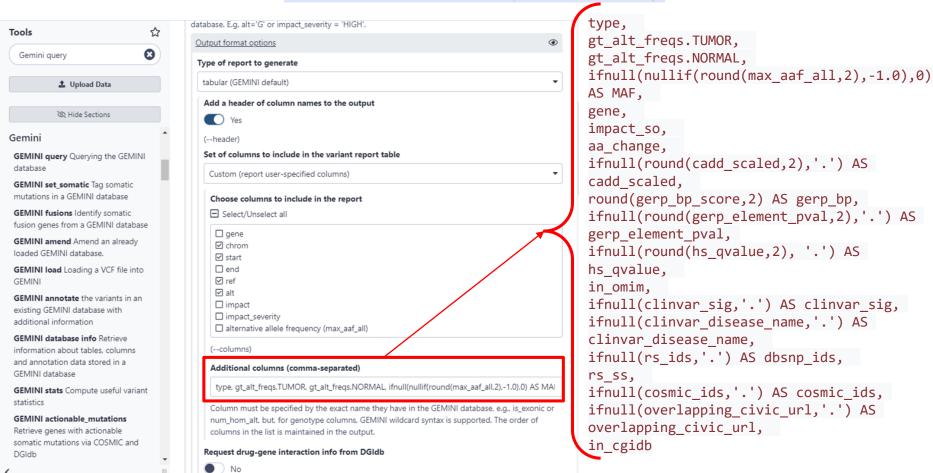
# b. Reporting selected subsets of variants

# Querying the GEMINI database for somatic variants



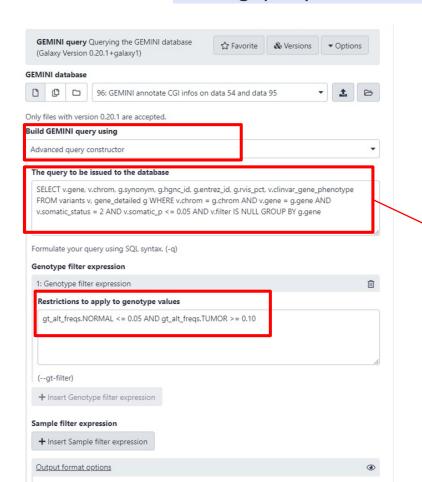


# **GEMINI SQL-based output formatting**



# c. Generating reports of genes affected by variants

# Turning query results into gene-centered reports



SELECT v.gene, v.chrom,
g.synonym, g.hgnc\_id,
g.entrez\_id, g.rvis\_pct,
v.clinvar gene phenotype

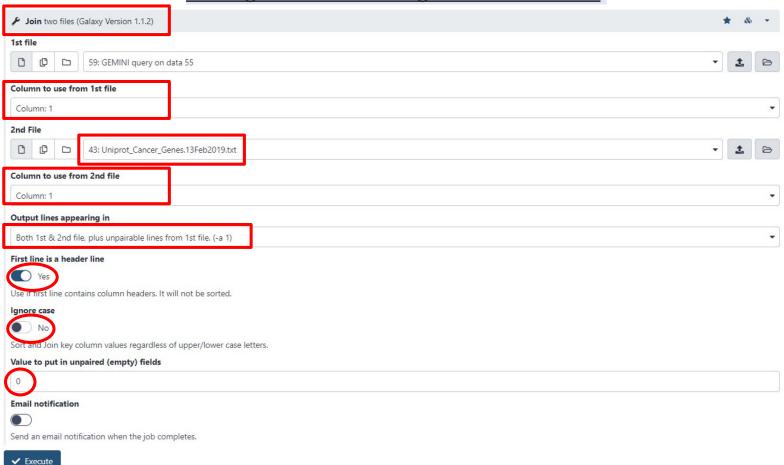
FROM variants v,
gene\_detailed g

WHERE v.chrom = g.chrom AND
v.gene = g.gene AND
v.somatic\_status = 2 AND
v.somatic\_p <= 0.05 AND
v.filter IS NULL</pre>

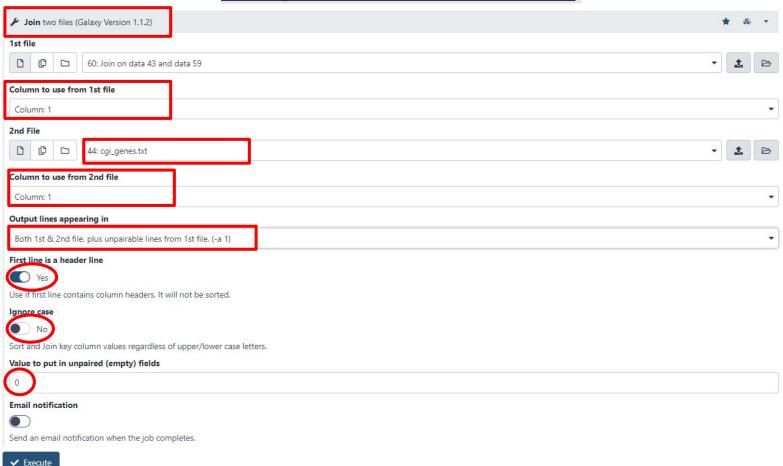
GROUP BY g.gene

# d. Adding additional annotations to the gene-centered report

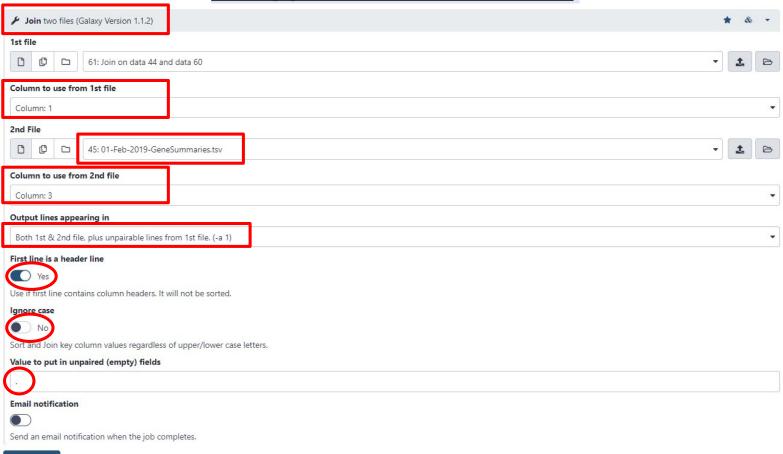
# Adding UniProt cancer genes information



# Adding CGI biomarkers information

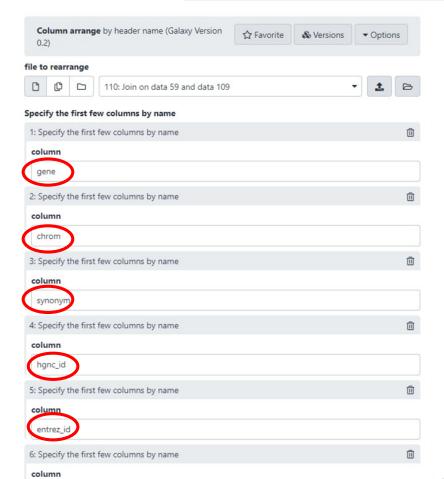


# Adding gene information from CIViC



✓ Execute

# Rearrange to get a fully annotated gene report



6: Specify the first few columns by name	Û
column	
rvis_pct	
7: Specify the first few columns by name	Û
column	
is_OG	
8: Specify the first few columns by name	Û
column	
is_TS	
9: Specify the first few columns by name	Û
column	
in_cgi_biomarkers	
10: Specify the first few columns by name	Û
column	
clinvar_gene_phenotype	
11: Specify the first few columns by name	Û
column	
gene_civic_url	
12: Specify the first few columns by name	Û
column	
description	
+ Insert Specify the first few columns by name	