# 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

### Starting from BAMs : Import Shared History

#### https://usegalaxy.fr/u/eag/h/somatic-tutorial

<b>= Galaxy</b> France			<b>☆</b>	Workflow	Visualize	Shared Data - Hel	lp 🕶	Log in or Register		4		
Tools search tools	☆ <b>*</b> <b>* ×</b>	Published Histories Search by name or use the advanced	d filter	ing options		Data Libraries Histories Workflows						* ×
1 Upload Data						Visualizations						Last
Get Data	•	Name	₽	Annotation	Owner	Pages		Tags				Updated 👻
Send Data		bilille_TP2_GTN_Somatic_Variants			eag							6 minutes
<b>Collection Operations</b>												ago
GENERAL TEXT TOOLS		imported: W4M00001_Sacurine-			julien.s	aint-vanne		age	bmi	gend	er	about 4
Text Manipulation		statistics						homos 3 more	sapiens	lcn	ns	hours ago
Filter and Sort								5 11016				
Join, Subtract and Group		coffee published 2022			amont	is						about 6
GENOMIC FILE MANIPULATION	I											hours ago
Convert Formats		Data_W4E_2024			cedric.	delporte						about 6
FASTA/FASTQ												hours ago
FASTQ Quality Control		TP W4E 2024			cdalle							about 7
SAM/BAM												hours ago

### Starting from BAMs : Import Shared History

#### https://usegalaxy.fr/u/eag/h/somatic-tutorial

\Xi Galaxy	France	*	Workflow V	isualize	Shared Data -	Help 🔻	Log in or Register	1	4	
Import this history										
search datasets	s									* ×
bilille_TP2_	_GTN_Somatic_Variants									/
🛢 5.11 GB										<b>Q</b> 10 <b>1</b> 2
12: mapped read	ds normal BAM									0
#normal	de tumor RAM									
#tumor										<b>v</b>
10: Uniprot_Cano	cer_Genes.13Feb2019.txt									Ø
9: sorted.correct	ted.01-Feb-2019-CIVic.bed									0
8: cgi_genes.txt	:									0
7: 01-Feb-2019-0	GeneSummaries.tsv									0
6: dbsnp.b147.cl	hr5_12_17.vcf.gz									Θ
5: 01-Feb-2019-0	CIVic.bed									Θ
4: cgi_variant_po	ositions.bed									0
3: hotspots.bed										0

### Prepare Data

- Click on the dataset
- Click on Set 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

a. Filtering on mapped reads properties

### Filtering for mapping status and quality

📮 Galaxy Europe	Analyze Data Workflow Visualize - Shared Data - Help - User - 📄 🇮			Using 12%	
Tools	Filter BAM datasets on a variety of attributes (Galaxy Version 2.4.1)	History search datasets	5	≎+⊡¢ 00	
L Upload Data     W Hide Sections Filter and Sort	BAM dataset(s) to filter 75: RmDup on data 73 74: RmDup on data 71 73: Filter on data 32: Filtered BAM 71: Filter on data 31: Filtered RAM 32: Map with BWA-MEM on tumor	Tumor Normal somatic pipeli 73 shown, 41 delete 31.3 GB	<b>l pair</b> ne TEST d, hide hido	den 🗹 📎 🗩	,
Column arrange by header name Filter data on any column using simple expressions Filter GTF data by attribute	31: Map with BWA-MEM on normal	tumor 76: BamLeftAlign a 74 (alignments)	) on dat	⊕ # ×	•
rate of the sequences by ID from a tabular file	Condition 1: Condition Filter	normal 75: RmDup on da tumor	ıta 73	④ ∦ ×	
Text Manipulation Add input name as column to an existing tabular file Sort Column Order by beading	1: Filter III Select BAM property to filter on	74: RmDup on da normal 73: Filter on data	ata 71 32: Filt	• / ×	
Replace column by values which are defined in a convert file Replace Text in a specific column	Filter on read mapping quality (phred scale) >=1 Yes capute a comparison Eq. to select reads with mapping quality of at	ered BAM tumor 72: Filter on data ON filter rules	32: JS	⊛ # ×	
Replace chromosome names in a tabular dataset using a mapping table Histogram of a numeric column Add column to an existing dataset	2: Filter "Insert Filter"     In "3: Filter":       Select BAM property to filter on     "Select DAM property to filter on	tumor	81: Filt	⊕ # ×	
Add column to an existing dataset Join two files on column allowing a small difference	isMapped         Selected mapped reads         Selected mapped reads         Select reads         Select reads	to filter vith	31: JS	⊕ # ×	
Add line to file writes a line of text at the begining or end of a text file.	Checked = Mapped, Empty = NOT mapped mate":	Yes	ieneS	<ul> <li></li></ul>	•

### 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

(-q)

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

Filter SAM or BAM, output SAM or BAM file by region (Galaxy Version 1.8+galaxy1)	es on FLAG MAPQ RG LN or 🏠 Favorite 🖓 Versions 💌 Options
SAM or BAM file to filter	
0         0         87: Filter on data 79: Filtered	I BAM - L
Header in output	
Include header	Skip alignments with any of these flag bits set
Minimum MAPQ quality score	Select/Unselect all
1	<ul> <li>Read is paired</li> <li>Read is mapped in a proper pair</li> </ul>
(-q)	<ul> <li>✓ The read is unmapped</li> <li>✓ The mate is unmapped</li> </ul>
	Read is mapped to the reverse strand of the reference
	Mate is mapped to the reverse strand of the reference
	Read is the first in a pair
	Read is the second in a pair
	The alignment of this read is not primary
	The read fails platform/vendor quality checks
	The read is a PCR or optical duplicate
	Supplementary alignment

## Mapped reads postprocessing

b. Removing duplicate reads

### Remove duplicates with RmDup

📮 Galaxy Europe	Analyze Data Workflow Visualize • Shared Data • Help • User • 😭 🏢	Using 12%
Tools ☆	RmDup remove PCR duplicates (Galaxy Version	Ĥistory ♂+□
Column arrange by header name 😢	2.0.1)	search datasets
🏝 Upload Data	BAM File           C         C         75: RmDup on data 73         L         C           Z4: RmDup on data 71         A         C         C         C	Tumor Normal pair somatic pipeline TEST
R Hide Sections	73: Filter on data 32: Filtered BAM 71: Filter on data 31: Filtered BAM	31.3 GB
Column arrange by header name	31 Map with BWA MEM on normal	
Filter data on any column using simple expressions	This is a batch mode input field. Separate jobs will be triggered for each dataset election.	76: BamLeftAlign on dat 🔹 🖋 🗙
Filter GTF data by attribute values_list	Is this paired-end or single end data	a 74 (alignments)
Filter sequences by ID from a tabular file	BAM is paired-end	75: RmDup on data 73 ④ 🖋 🗙
Text Manipulation	No	74: Per Dun an data 71 (A A Y
Add input name as column to an existing tabular file	(-S)	normal
Sort Column Order by heading	Email notification	format: <b>bam</b> , database: <b>hg19</b>
Replace column by values which are defined in a convert file	No Send an email notification when the iob completes.	[bam_rmdup_core] processing
Replace Text in a specific column		[bam_rmdup_core] 1 unmatched pairs
Replace chromosome names in a tabular dataset using a mapping table	✓ Execute	[bam_rmdup_core] processing reference chr11
Histogram of a numeric column	What it does	[bam_rmdup_core] 3 unmatched pairs [bam_rmdup_core] processing
Add column to an existing dataset	Remove potential PCR duplicates: if multiple read pairs have identical external coordinates, only retain the	reference chr12
Add column to an existing dataset	pair with highest mapping quality. In the paired-end mode, this command ONLY works with FR orientation and requires ISIZE is correctly set. It does not work for unpaired reads (e.g. two ends mapped to different	[bam_rmdup_core] inconsistent BAM file for
Join two files on column allowing a small difference	chromosomes or orphan reads).	80082 <b>₩</b> ? ♥●
Column Regex Find And Replace	Citations II	display at UCSC main
Add line to file writes a line of text at the begining or end of a text file.	- Definition of SAM/BAM format. (n.d.). Retrieved from https://samtools.github.io/hts-specs/ 🔀 - Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., and, R. D. (2009). The Sequence	display at Ensembl Current display with IGV local Human hg19

\* >

## Mapped reads postprocessing

c. Left-align reads around indels

### Left-align with BamLeftAlign

📱 Galaxy Europe	Analyze Data 🛛 Workflow Visualize = Shared Data = Help = User = 📷	•	Using 91	<b>%</b>
Tools 🗘	BamLeftAlign indels in BAM datasets (Galaxy Version 1.3.1)	His	itory 🕄 + 🗖	•
BamLeftAlign		Added to versions Coptions	earch datasets	Θ
2. Upload Data	Choose the source for the reference genome Locally cached	Tur • Sor	nor Normal pair natic pipeline TEST	
R Hide Sections	Sel <mark>e crangment</mark> file in BAM format	32 s	hown, 34 deleted, 13 hidden	
VCF/BCF	C C 75: RmDup on data 73	• 1 🖻 228	.7 GB ⊠ ♥	
BamLeftAlign indels in BAM datasets	Using reference genome	76:	BamLeftAlign on da 🛛 👁 🖋 🗙	e ^
Variant Calling	Human (Homo sapiens): hg19	- ta 7	4 (alignments)	
BamLettAlign indels in BAM datasets	(fasta-reference)	form	nat: bam, database: hg19	
All workflows	Maximum number of iterations	3	∂02≝? <b>●</b> ●	2
	5 Iterate the left-realignment no more than this many times (max-iterations) Email notification  No	disp disp disg disg disp	lay at UCSC main Ilay at Ensembl Current Ilay with IGV local Human hg19 Ilay in IGB View Ilay at bam.iobio bam.iobio.io	
	Send an email notification when the job completes.	76.		
	✓ Execute	13.		
	When calling indels, it is important to homogenize the positional distribution of insertions and deletions in the input by using left realignment. Left realignment wi provided that doing so does not introduce mismatches between the read and reference other than the indel. This method is computationally inexpensive and han	ill place all indels in homopolymer and microsatellite repeats at the same position, dles the most common classes of alignment inconsistency.	Filter on data 32: Fil ③ / × ed BAM	¢
	This is leftalign utility from FreeBayes package.	72: ON	Filter on data 32: J5 ④ 🖋 🗙 filter rules	r.
	Citations: - (N.d.). Retrieved from http://anxiv.org/abs/1207.3907	71: tere nor	Filter on data 31: Fil ④ 🖋 🗙 ed BAM mal	2
	- freebayes (Version 1.3.1) - samtools (Version 1.9)	70: ON	Filter on data 31: JS ④ 🖋 🗙 filter rules mal	¢.
		32: M c Tur	Map with BWA-ME 🕢 🖋 🗙 on tumor not	8
		31: M c	Map with BWA-ME 🔹 🖋 🗙 on normal	¢ +
<	Ē.	1		>

## Mapped reads postprocessing

d. Recalibrate read mapping qualities

### Recalibrate read quality scores with CalMD

💶 Galaxy Europe	Analyze Data Workflow Visualize * Shared Data * Help * User * 🞓 🏢	Using 13%
Tools t	CaIMD recalculate MD/NM tags (Galaxy Version 2.0.2)	Ĥistory ♂+⊡≮
filter 😣	Added & Versions + Options	search datasets
오 Hide Sections	BAM file to recalculate          T7: BamLeftAlign on data 75 (alignments)         76: BamLeftAlign on data 74 (alignments)         75: RmDup on data 73	Tumor Normal pair somatic Tutorial 40 shown, 45 deleted, 29 hidden
Sharpen	74: RmDup on data 71 73: Filter on data 32: Filtered BAM	31.3 GB 🗹 🍽 🗩
deepTools estimateReadFiltering estimates the number of reads that would be	71: Filter on data 31: Filtered BAM 32: Man with BWA-MEM on tumor This is a batch mode input field. Separate jobs will be triggered for each dataset selection. Choose the source for the reference genome	84: Filter on data 78: JS ④ 🖋 🗙 ON filter rules normal
filtered given certain criteria alignmentsieve Filter BAM/CRAM files according to specified	Use a built-in genome	79: CaIMD on data 77 ③ 🖋 🗙
parameters SAM/BAM	Human (Homo sapiens): hg19	78: CalMD on data 76 💿 🥒 🗙
Filter BAM datasets on a variety of attributes	Do you also want BAQ (Base Alignment Quality) scores to be calculated?	754.4 MB format: <b>bam</b> , database: <b>hg19</b>
BAM filter Removes reads from a BAM file based on criteria	(-r)	[bam_fillmd1] different NM for read 'ST-
Samtools view - reformat, filter, or subsample SAM, BAM or CRAM	Additional options	0 -> 1 [bam_fillmd1] different MD for read
Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region	Change identical bases to '='	'ST- K00265:137:HT33CBBXX:3:2115:18345: '98' -> '33n64'
MiModD	Replace bases in read sequences that match the reference base at that position with an equal sign (-e)	[bam_fillmd1] different MD for read 'ST-K00265:137:HT33CBBXX:3:1127
MiModD VCF Filter extracts lines from a vcf variant file based on field- specific filters	Coefficient to cap mapping quality of poorly mapped reads       50	<
Metagenomic Analysis	righer values for this setting mean a stronger downgrade of the mapping quality of reads with excessive mismatches (50:	display at UCSC main
dada2: filterAndTrim Filter and trim short read data	recommended setting for reads aligned with BWA, U: do not downgrade mapping qualities) (-C) Email notification	display with IGV local Human hg19 display in IGB View
sixgill filter a metapeptide database	No No	display at bam.iobio bam.iobio.io
khmer: Filter reads by minimal k-	Send an email notification when the job completes.	v

## Mapped reads postprocessing

e. Refilter reads based on mapping quality

### Eliminating reads with undefined mapping quality

📮 Galaxy Europe	Analyze Data 🛛 Workflow 🖓 Visualize * Shared Data * Help * User * 📂 🇮	
Tools 🗘	Filter BAM datasets on a variety of attributes (Galaxy Version 2.4.)	
Filter BAM		Addeo ex versions Options
2, Upload Data	BA reasonable to filter	- <b>1</b> B
R Hide Sections           overlapping/reality merivals           bedtools Genome Coverage           compute the coverage over an entire           genome	7: BamLeftAlign on data 75 (alignments) 7: BamLeftAlign on data 74 (alignments) 7: RmDup on data 71 74: RmDup on data 71 72: Eliterand tata 27: Eliterand RAM This is a batch mode input field. Separate jobs will be triggered for each dataset selection.	3
SAM/BAM	Condition	
BedCov calculate read depth for a set of genomic intervals	1: Condition	
BAM filter Removes reads from a BAM file based on criteria	Filter 1: Filter	
Convert, Merge, Randomize BAM datasets and perform other transformations	Select BAM property to filter on mapQuality	-
BamHash Hash BAM and FASTQ files to verify data integrity	Filter on read mapping quality (phred scale) <=254	
attributes	You can use >, <, =, and ! (not) in your expression. E.g., to select reads with mapping quality of at least 30 use ">= 30"	
CaIMD recalculate MD/NM tags	+ Insert Filter	
FASTA/FASTQ	+ Insert Condition	
Create binary barcodes from regular barcodes.	Would you like to set rules?	
Extract barcodes according to pattern	No	
Barcode Splitter	Allows complex logical constructs. See Example 4 below.	
UMI-tools whitelist Extract cell barcodes from FASTQ files	No	
Filter sequences by mapping from SAM/BAM file	Send an email notification when the job completes.	
Je-Demultiplex demultiplexes fastq files	✓ Execute	
Extract alignment ends from SAM or BAM	What is does BAMTools filter is a very powerful utility to perform complex filtering of BAM files. It is based on BAMtools suite of tools by Derek Barnett (https://github.com/pezmaster31/bamtools)	
Proteomics	r Haw it works	

## 2. Variant calling and classification

### Variant calling with VarScan somatic

🚆 Galaxy Europe	Analyze Data 🛛 Workflow Visualize * Shared Data * Help * User * 📻 🏥	
Tools 🗘 VarScan 😵	VarScan somatic Call germline/somatic and LOH variants from tumor-normal sample pairs (Galaxy Version 2.4.3.6)	★ Added & Versions ♥ Options
	Will you select a reference genome from your history or use a built-in genome?	
2. Upload Data	Use a built-in genome	•
R Hide Sections	reference genome	
VCF/BCF	Human (Homo sapiens): hg19	•
VarScan somatic Call germline/somatic	The fasta reference genome that variants should be called against.	
and LOH variants from tumor-normal sample pairs	aligned reads from normal sample	
VarScan mpileup for variant detection	D D B7: Filter on data 79: Filtered BAM	- 1 0
VarScan copynumber Determine	aligned reads from tumor sample	
relative tumor copy number from tumor- normal pileups	0 0 87: Filter on data 79: Filtered BAM	• 1 B
VarScan for variant detection		
Variant Calling	Estimated purity (non-tumor content) of normal sample	
VarScan mpileup for variant detection		
VarScan copynumber Determine	(normal-purity)	
normal pileups	Estimated purity (tumor content) or tumor sample	
VarScan somatic Call germline/somatic	0.5	
and LOH variants from tumor-normal sample pairs	(tumor-punity) Generate construct datasets for SNP and indel calle?	
WORKFLOWS		
All workflows	Settings for Variant Calling	
	Use default values	•
	Settings for Posterior Variant Filtering	
		•
	Compatibility options for experts	Ø
	Email notification	
	• No	
	Send an email notification when the job completes.	
	✓ Execute	
<	Ver-Fran Romanian	

### Variant calling with VarScan somatic

🚆 Galaxy Europe	Analyze Data 🛛 Workflow Visualize * Shared Data * Help * User * 🎓 🇮
Tools VarScan	VarScan somatic Call germline/somatic and LOH variants from tumor-normal sample pairs (Galaxy Version 2.4.3.6)
	Will you select a reference genome from your history or use a built-in genome?
🍰 Upload Data	Use a built-in genome
े Hide Sections	reference genome
VCF/BCF	Human (Homo sapiens): hg19
VarScan somatic Call germline/somati	The fasta reference genome that variants should be called against.
and LOH variants from tumor-normal	aligned reads from normal sample
VarScan mpileup for variant detection	D D 87: Filter on data 79: Filtered BAM
VarScan copynumber Determine relative tumor copy number from tumo normal pileups	r- D D 87: Filter on data 79: Filtered BAM
VarScan for variant detection	
Variant Calling	Estimated purity (non-tumor content) of normal sample
VarScan mpileup for variant detection	1
VarScan copynumber Determine relative tumor copy number from tumo normal pileups	(normal-purity) r- Estimated purity (tumor content) of tumor sample
VarScan somatic Call germline/somati and LOH variants from tumor-normal sample pairs	(tumor-purity) Settings for Variant Calling
WORKFLOWS	Generate separate output datasets for SNP and indel calls? Customize settings
All workflows	Settings for Variant Calling
	Kead selection
	Setting for Partners United Elleving
	28
	Compatibility options for experts molecup - Q)
	Email notification Minimum mapping quality
	• No
	Send an email notification when the job completes.
	The minimum mapping quality (default: 0) required for a read to be considered in variant calling (samtools mpileup -q)

VarScan Overview

<

## 3. Variant annotation and reporting

### Adding annotations to the called variants

### a. Adding annotations to the called variants

a.1. Adding functional genomic annotations

### Adding annotations with SnpEff

📮 Galaxy Europe		Analyze Data Workflow Visualize - S	hared Data + Help + I	Jser - 😥					
Tools SnpEff eff	1	SnpEff eff: annotate variants (Galaxy Version 4.3+T.galaxy1)					☆ Favorite 🔹 Ve	rsions 🔹 Op	otions
🛓 Upload Data		Sequence changes (SNPs, MNPs, InDels)           Image: Display the sequence of th						- 1	<b>t</b> B
🗞 Hide Sections		Input format							
VCF/BCF		VCF							•
<b>snippy</b> Snippy finds SNPs between a haploid reference genome and your NGS sequence reads.		Output format VCF (only if input is VCF)							•
snippy-core Combine multiple Snippy outputs into a core SNP alignment		Create CSV report, useful for downstream analysis (-csvStats)							
ococo consensus caller on SAM/BAM		Genome source							
Variant Calling		Locally installed snpEff database							•
snippy-core Combine multiple Snippy outputs into a core SNP alignment		Genome Homo sapiens : hg19							•
snippy-clean_full_aln Replace any non-standard sequence characters in spinny 'rore full aln' file		Regulation options							Ø
SnpEff eff: annotate variants for SARS-CoV-2		Upstream / Downstream length 5000 bases							•
SnpEff eff: annotate variants		(-ud)							
SnpEff databases: list available		Set size for splice sites (donor and acceptor) in bases							
databases		2 bases							•
built database		(-55)							
Variant Frequency Plot Generates a		spliceRegion Settings							
heatmap of allele frequencies grouped by variant type for SnpEff-		Use Defaults			<b>Г</b>	Due due a Cur			
annotated SARS-CoV-2 data		Annotation options				Floauce Sul	illiary stats		
SnpEff chromosome-info: list chromosome names/lengths		Select/Unselect all				No No			
SnpEff build: database from Genbank or GFF record		Use 'EFF' field compatible with older versions (instead of 'ANN') Use Classic Effect names and amino acid variant annotations (NON_SYNONYMOUS_CODING vs missense_variant ann Override classic and use Sequence Ontolgy terms for effects (missense_variant vs NON_SYNONYMOUS_CODING)	i G180R vs p.Gly180Arg/c	.538G>C)		(-noStats)			
SnpEff to Peptide fasta to create a Search DB fasta for variant SAP peptides	•	Override classic and use HGVS annotations for amino acid annotations (p.Gly180Arg/c.538G>C vs G180R)     Old notation style notation: Eg. 'C.G123T instead of ".123G>T and X' instead of "*'     Use one letter Amino acid codes in HGVS notation. Eg. p.RA7G instead of p.Arg47Gly     Use transcript ID in HGVS notation. Eg. ENST00000252100:c914C>G instead of c.914C>G							

### a. Adding annotations to the called variants

a.2. Adding genetic and clinical evidence-based annotations

### Creating a GEMINI database from a variants dataset

🗧 Galaxy Europe	Analyze Data 🛛 Workflow Visualize 🕆 Shared Data 🔨 Help 🎽 User 🏲 📂 🇮		Using 8%
Tools 🖒		History	2+□¢
GEMINI load	GEMINI load Loading a VCF file into GEMINI (Galaxy Version 0.20.1+galaxy2)	search datasets	00
🏦 Upload Data	VCF dataset to be loaded in the GEMINI database	Tumor Normal pair somatic pipeline TEST	г
জ্ঞ Hide Sections	Only by ild 37 (aka bo19) of the human genome is supported.	41 shown, 39 deleted, 13 hidde	en en en
RNA Analysis	The variants in this input are	21.27 GB	
StringTie merge transcripts	annotated with snpEff	90: SnoEff.eff: on data 8	@ # X
Gene Body Coverage (Bigwig) Read	GEMINI can parse and use annotations generated with either snpEff (both 'EFF' and 'ANN'-style annotations are supported) or VEP. You can also load unannotated variants, but most of GEMINI's functionality will not be available or not be very useful	8	
StringTie transcript assembly and	This input comes with genotype calls for its samples	88: VarScan somatic on	⊕∤×
quantification	Ves	data 87 and data 85	
Gemini	This is usually the case, but some published datasets, like some 1000G VCFs, are missing genotype information. (no-genotypes)	87: Filter on data 79: Fil tered BAM	⊕ # ×
GEMINI set_somatic Tag somatic mutations in a GEMINI database	Choose a gemini annotation source	86: Filter on data 79: JS	@ / X
GEMINI load Loading a VCF file into	GEMINI annotations w/ GERP & CADD (2019-01-12 snapshot)	ON filter rules	
GEMINI	Sample and family information in PED format	85: Filter on data 78: Fil	@ / X
GEMINI fusions Identify somatic fusion genes from a GEMINI database	C 🖸 Nothing selected	tered BAM	
GEMINI amend Amend an already loaded GEMINI database.	The pedigree dataset is optional, but several GEMINI tools require the relationship between samples (i.e., the family structure) and/or the sample phenotype to be defined. The PED format is a simple tabular format (see the tool help below for	84: Filter on data 78: JS ON filter rules	⊕ ∦ X
GEMINI query Querying the GEMINI	Load the following optional content into the database	79: CalMD on data 77	
database	E Select/Unselect all	tumor	
GEMINI annotate the variants in an existing GEMINI database with	C GERP scores	78: CalMD on data 76	@ / X
additional information	CADD scores (non-commercial use only; see licensing note below)	normal	
GEMINI database info Retrieve	⊘ sene tables ② Sample genotypes	77: BamLeftAlign on da	⊕ # ×
and annotation data stored in a	Genotype likelihoods (sample PLs)	ta /5 (alignments)	
GEMINI database	only variants that passed all initiars     original variants in the passed all initiars     original variant INFO field		
GEMINI stats Compute useful variant	The exercise of defaulty cloud has deferences the second to a part the second to a part to a part of the second to detail the second to	76: BamLeftAlign on da ta 74 (alignments)	• / ×
statistics	me preserved versions make an use or normous bases (see new or mour base or source and or mour base) and the preserved versions and the preserved versions and the preserved version and versions and ve	normal	
Retrieve genes with actionable	Email notification	75: RmDup on data 73	@ / X
somatic mutations via COSMIC and	No	tumor	
GEMINI hurden national samela	Send an email notification when the job completes.	74: RmDup on data 71	@ / X
wise gene-level burden calculations		normal	
GEMINI rob Identifying runs of		73: Filter on data 32: Fil	@#¥

cr					History	<b>₽+</b> □
(Ga	laxy Version	0.20.1+galaxy2)	ions • Op	tions	search datasets	00
EMI	VI database				Tumor Normal pair	
٥	¢ 🗅	91: GEMINI load on data 90	t -	. 0	somatic pipeline TES	ST
nly fi	les with vers	ion 0.20.1 are accepted.			42 shown, 39 deleted, 13 hid 30.2 GB	iden
atas	et to use as	the annotation source				
0	m Ch	20. Vectors constitute of data 07 and data 05				
	<u> </u>	oc: Varscan somatic on data or and data op	•	. 6	91: GEMINI load on dat a 90	● # ×
he to trict	ol can use th	ne information from a BED or VCF dataset to annotate the database variants. (-f) The information of database and annotation records (VCF format only)			91: GEMINI load on dat a 90 normal tumor	● # ×
he to trict	ol can use th variant-ider Yes	ne information from a BED or VCF dataset to annotate the database variants. (-f)			91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8	• / ×
he to trict	ol can use th variant-ider Yes	be: varscan somatic on data s7 and data s5 ne information from a BED or VCF dataset to annotate the database variants. (-f) ntity matching of database and annotation records (VCF format only)	tion source des	scribe the	91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8 normal tumor	● # ×
he to trict he de xact s	ol can use th variant-ider Yes fault is to co same nucleo sition of a d	be: varscan somatic on data 87 and data 85 ne information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> ponsider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annota tide change at the same position in the genome. You can disable this option to make use of any annot atabase variant. This setting is ignored for annotation sources in BED format, for which matching is alw	tion source des ation that over ays based on	cribe the laps with	91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8 normal tumor 88: VarScan somatic on data 97 and data 95	● # × ● # ×
he to trict he de xact s he po verla	ol can use th variant-ider Yes efault is to cco same nucleo sition of a d ppping positio	be: varscan somatic on data s7 and data s5 me information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> ponsider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annota tide change at the same position in the genome. You can disable this option to make use of any annot atabase variant. This setting is ignored for annotation sources in BED format, for which matching is alw ons only. (region-only)	tion source des ation that over ays based on	cribe the laps with	91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8 normal tumor 88: VarScan somatic on data 87 and data 85 normal tumor	● # × ● # ×
he to trict he de xact s ne po verla	ol can use th variant-ider Yes fault is to cc ame nucleo sition of a d pping positic of informati	be information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> ponsider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annota tide change at the same position in the genome. You can disable this option to make use of any annot atabase variant. This setting is ignored for annotation sources in BED format, for which matching is alw ons only. (region-only) <b>fon to add to the database variants</b>	tion source des ation that over ays based on	cribe the laps with	91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8 normal tumor 88: VarScan somatic on data 87 and data 85 normal tumor	● # × ● # × ● # ×
he to trict he de kact s verla ype o Spec	ol can use th variant-ider Yes afault is to cco same nucleo sition of a d pping positio of informati ific values ex	be: varscan somatic on data s7 and data s5 the information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> ponsider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annota tide change at the same position in the genome. You can disable this option to make use of any annot atabase variant. This setting is ignored for annotation sources in BED format, for which matching is alw ons only. (region-only) tion to add to the database variants ctracted from matching records in the annotation source (extract)	tion source des ation that over ays based on	scribe the laps with	91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8 normal tumor 88: VarScan somatic on data 87 and data 85 normal tumor 87: Filter on data 79: Fil tered BAM	● # × ● # × ● # ×
ne to trict ne de xact s ve po verla ype o Spec a)	ol can use th variant-ider Yes fault is to co ame nucleo sition of a d pping positio of informati ific values ex	be information from a BED or VCF dataset to annotate the database variants. (-f) <b>Initiy matching of database and annotation records (VCF format only)</b> Donsider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annotatide change at the same position in the genome. You can disable this option to make use of any annot atabase variant. This setting is ignored for annotation sources in BED format, for which matching is alw ons only. (region-only) <b>fon to add to the database variants tracted</b> from matching records in the annotation source (extract)	tion source destation that over	cribe the laps with	91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8 normal tumor 88: VarScan somatic on data 87 and data 85 normal tumor 87: Filter on data 79: Fil tered BAM tumor	● # × ● # × ● # ×
e to rict e de act : e po erla pe ( bpec	ol can use th variant-ider Yes efault is to co ame nucleo sition of a d. pping position of informati ific values ex otation extr	be: varscan somatic on data s7 and data s5 me information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> ponsider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annota tide change at the same position in the genome. You can disable this option to make use of any annot atabase variant. This setting is ignored for annotation sources in BED format, for which matching is alw ons only. (region-only) <b>fon to add to the database variants</b> stracted from matching records in the annotation source (extract) <b>raction recipe</b>	tion source des ation that over ays based on	cribe the laps with	91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8 normal tumor 88: VarScan somatic on data 87 and data 85 normal tumor 87: Filter on data 79: Fil tered BAM tumor 86: Filter on data 79: JS	● # × ● # × ● # × ● # ×
he to trict he de kact s he po verla ype o Spec a) Ann 1: 4	ol can use th variant-ider Yes efault is to co same nucleo sition of a d pping positio of informati ific values ex otation extr annotation e	be: varscan somatic on data s7 and data s5 me information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> ponsider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annota tide change at the same position in the genome. You can disable this option to make use of any annot atabase variant. This setting is ignored for annotation sources in BED format, for which matching is alw ons only. (region-only) <b>for to add to the database variants</b> ctracted from matching records in the annotation source (extract) <b>raction recipe</b> xtraction recipe	tion source destation that over	cribe the laps with	91: GEMINI load on dat a 90 normal tumor 90: SnpEff eff: on data 8 8 normal tumor 88: VarScan somatic on data 87 and data 85 normal tumor 87: Filter on data 79: Fil tered BAM tumor 86: Filter on data 79: JS ON filter rules	• / × • / × • / ×

Analyze Data Workflow Visualize - Shared Data - Help - User - 📂 🌉			Using 12%
1: Annotation extraction recipe	1	History	2+04
Elements to extract from the annotation source		search datasets	00
SS 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_status		Tumor Normal pair somatic pipeline TES 42 shown, 39 deleted, 13 hid 30.2 GB	den
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		91: GEMINI load on dat a 90 normal tumor	● / ×
<ul> <li>O Integer numbers</li> <li>O Text (text)</li> <li>Your selection will determine the data type used to store the new annotations in the database. (-t)</li> </ul>		90: SnpEff eff: on data 8 8 normal tumor	⊕ # X
If multiple annotations are found for the same variant, store		88: VarScan somatic on	• / ×
the first annotation found		data 87 and data 85	
Note: If indicated (in parentheses) an option is only applicable to annotations of a specific type. (-o)		97: Filher and data 70: Fil	
2: Annotation extraction recipe		tered BAM	
Elements to extract from the annotation source		tumor	
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,	]	86: Filter on data 79: JS ON filter rules tumor	<b>④</b> ∦ ×
name an invro neig element. (-e)	*	1	

Note: If indicated (in parentheses) an option is only applicable to annotations of a specific type, (-o)	* u	istory	0 + m 4
2: Annotation extraction recipe		istory	
Elements to extract from the annotation source		search datasets	00
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)	Tr 50 42	umor Normal pair omatic pipeline TEST 2 shown, 39 deleted, 13 hidde	:n
		72 GB	
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?	91 a	1: GEMINI load on dat 90 rormal tumor	● / ×
<ul> <li>Numbers with decimal precision</li> <li>Integer numbers</li> <li>Text (text)</li> </ul>	90 8	D: SnpEff eff: on data 8	⊛ / ×
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	88	8: VarScan somatic on ata 87 and data 85	● # ×
the first annotation found		iormai tumor	-
Note: If indicated (in parentheses) an option is only applicable to annotations of a specific type. (-o)	87 te	7: Filter on data 79: Fil red BAM	• / ×
3: Annotation extraction recipe	J T	umor	
Elements to extract from the annotation source	86 Ol	5: Filter on data 79: JS N filter rules umor	● / ×
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)	+		

Annotation extraction recipe	1	History	<b>2 + D</b>
lements to extract from the annotation source		search datasets	00
SPV or an annotation source in BED format, specify the number of the column from which the annotations should be read. For a VCF source, ame an INFO field element. (-e) Tatabase column name to use for recording annotations		Tumor Normal pair somatic pipeline TES 42 shown, 39 deleted, 13 hidd 30.2 GB	r Ien
somatic_p column with the name provided here will be added to the variants table of the GEMINI database to store the annotations (-c)		91: GEMINI load on dat a 90	⊕ # ×
O Numbers with decimal precision O Integer numbers O Text (text)		normal tumor 90: SnpEff eff: on data 8 8 normal tumor	⊛ # ×
our selection will determine the data type used to store the new annotations in the database. (-t) <b>multiple annotations are found for the same variant, store</b> the first annotation found		88: VarScan somatic on data 87 and data 85 normal tumor	⊙ / ×
lote: If indicated (in parentheses) an option is only applicable to annotations of a specific type. (-o) Insert Annotation extraction recipe		87: Filter on data 79: Fil tered BAM tumor	⊕ # ×
I notification		86: Filter on data 79: JS ON filter rules	⊛ / ×

### Adding further annotations from dbSNP

CEMINI			Î	History	2+0
(Galary Version	Cate the variants in an existing GEMINI database with additional information C 20.1 - Selevy3	♥ Options		search datasets	0
EMINI database	92: GEMINI annotate on data 88 and data 91	• <u>t</u>	2	Tumor Normal pair somatic pipeline TES	т
nly files with vers	sion 0.20.1 are accepted.			56 shown, 39 deleted, hide hi 30.2 GB	idden 🗹 🃎
000	56: dbsnp.b147.chr5_12_17.vcf.gz	t.	2	es.13Feb2019.txt	
					x
ne tool can use ti trict variant-ider	he information from a BED or VCF dataset to annotate the database variants. (-f) ntity matching of database and annotation records (VCF format only)			56: dbsnp.b147.chr5_12 _17.vcf.gz	
ne tool can use the trict variant-ider Yes Yes	he information from a BED or VCF dataset to annotate the database variants. (-f) ntity matching of database and annotation records (VCF format only)	urce describe	the	56: dbsnp.b147.chr5_12 _17.vcf.gz 55: 01-Feb-2019-CIVic.b ed	⊙ # ×
trict variant-ider Ves Ves vact same nucleo vert same nucleo vert same nucleo vert same nucleo	he information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> possider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annotation sou stide change at the same position in the genome. You can disable this option to make use of any annotation that latabase variant. This setting is ignored for annotation sources in BED format, for which matching is always base out one only (-region-only)	urce describe at overlaps w ed on	the ith	56: dbsnp.b147.chr5_12 _17.vcf.gz 55: 01-Feb-2019-CIVic.b ed 54: cgi_variant_position s.bed	@ / ×
Trict variant-ider version of a d verlapping positiv ver of informati	he information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> <u>onsider VCE-formatted annotations only if a variant in the GEMINI database</u> and a record in the annotation sou itide change at the same position in the genome. You can disable this option to make use of any annotation that latabase variant. This setting is ignored for annotation sources in BED format, for which matching is always base ons only. (region-only) <b>ion to add to the database variants</b>	urce describe at overlaps w ed on	the ith	56: dbsnp.b147.chr5_12 _17.vcf.gz 55: 01-Feb-2019-CIVic.b ed 54: cgi_variant_position s.bed 53: hotspots.bed	• / × • / × • / ×
Trict variant-ider trict variant-ider Ves be default is to co kact same nucleo be position of a d verlapping positive vpe of informations Specific values exp	he information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> consider VCE-formatted annotations only if a variant in the GEMINI database and a record in the annotation sou itide change at the same position in the genome. You can disable this option to make use of any annotation the latabase variant. This setting is ignored for annotation sources in BED format, for which matching is always base ons only. (region-only) <b>ion to add to the database variants</b> xtracted from matching records in the annotation source (extract)	urce describe at overlaps w ed on	the ith	56: dbsnp.b147.chr5_12 _17.vcf.gz 55: 01-Feb-2019-CIVic.b ed 54: cgi_variant_position s.bed 53: hotspots.bed 32: Map with BWA-ME	<ul> <li></li></ul>
Trict variant-ider trict variant-ider Ves be default is to control to cont	he information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> consider VCE-formatted annotations only if a variant in the GEMINI database and a record in the annotation sou itide change at the same position in the genome. You can disable this option to make use of any annotation the latabase variant. This setting is ignored for annotation sources in BED format, for which matching is always base ons only. (region-only) <b>ion to add to the database variants</b> xtracted from matching records in the annotation source (extract)	urce describe at overlaps w ed on	the ith	56: dbsnp.b147.chr5_12 _17.vcf.gz 55: 01-Feb-2019-CIVic.b ed 54: cgi_variant_position s.bed 53: hotspots.bed 32: Map with BWA-ME M on tumor	<ul> <li></li></ul>
Annotation extin	he information from a BED or VCF dataset to annotate the database variants. (-f) <b>ntity matching of database and annotation records (VCF format only)</b> possider VCE-formatted appotations only if a variant in the GEMINI database and a record in the annotation sou stide change at the same position in the genome. You can disable this option to make use of any annotation that latabase variant. This setting is ignored for annotation sources in BED format, for which matching is always base ons only. (region-only) <b>ion to add to the database variants</b> xtracted from matching records in the annotation source (extract) <b>raction recipe</b>	urce describe at overlaps w ed on	the ith	56: dbsnp.b147.chr5_12 17.vcf.gz 55: 01-Feb-2019-CIVic.b ed 54: cgi_variant_position s.bed 53: hotspots.bed 32: Map with BWA-ME M on tumor tumor	• / × • / × • / × • / ×
Annotation extension of a diversion	he information from a BED or VCF dataset to annotate the database variants. (-f) ntity matching of database and annotation records (VCF format only) posider VCE-formatted annotations only if a variant in the GEMINI database and a record in the annotation sou itide change at the same position in the genome. You can disable this option to make use of any annotation that latabase variant. This setting is ignored for annotation sources in BED format, for which matching is always base ons only. (region-only) ion to add to the database variants xtracted from matching records in the annotation source (extract) raction recipe	urce describe at overlaps w ed on	the ith	56: dbsnp.b147.chr5_12 17.vcf.gz 55: 01-Feb-2019-CIVic.b ed 54: cgi_variant_position s.bed 53: hotspots.bed 32: Map with BWA-ME M on tumor tumor 31: Map with BWA-ME M on normal	<ul> <li></li></ul>

### Adding further annotations from dbSNP

vact same nucleotide change at the same position in the genome. You can disable this option to make use of any annotation that overlaps with ne position of a database variant. This setting is ignored for annotation sources in BED format, for which matching is always based on	*	History	<b>₽+</b> □	\$
verlapping positions only. (region-only)		search datasets	00	3
pe of information to add to the database variants		C		2
Specific values extracted from matching records in the annotation source (extract) •		Tumor Normal pair somatic pipeline TES	т	
a) Annotation extraction recipe		56 shown, 39 deleted, hide h	idden	
1: Annotation extraction recipe		30.2 GB		•
Elements to extract from the annotation source		es.13Feb2019.txt		
For an annotation source in BED format, specify the number of the column from which the annotations should be read. For a VCF source,		56: dbsnp.b147.chr5_12 _17.vcf.gz	⊛ ∥ ×	
name an INFO field element. (-e) Database column name to use for recording annotations		55: 01-Feb-2019-CIVic.b ed	● / ×	
15_55		54: cgi_variant_position	• / ×	
A column with the name provided here will be added to the variants table of the GEMINI database to store the annotations (-c)		s.bed		
What type of data are you trying to extract?		53: hotspots.bed	⊙ / ×	
<ul> <li>Numbers with decimal precision</li> <li>Integer numbers</li> <li>Text (text)</li> </ul>	-	32: Map with BWA-ME M on tumor	⊛ # ×	
Your selection will determine the data type used to store the new annotations in the database. (-t)		tumor		
If multiple annotations are found for the same variant, store		31: Map with BWA-ME	⊙ / ×	
the first annotation found		normal		
Note: If indicated (in narentheses) an ontion is only applicable to apportations of a specific type (-o)	+	1		

### Adding further annotations from Cancer Hotspots v2

Analyze Data Workflow Visualize * Shared Data * Help * User * 📂 🏢			Using 12%
Dataset to use as the annotation source	-	History	ଟ+⊡¢
D         D         53: cancerhotspots_v2.bed         ▼         1		search datasets	00
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)  Ves	- 1	Tumor Normal pair somatic pipeline TES	51
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 option to 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)		30.42 GB	🗹 📎 🗩
Type of information to add to the database variants		93: GEMINI annotate o n data 56 and data 92	• # ×
(-a) Annotation extraction recipe		92: GEMINI annotate o n data 88 and data 91 normal tumor	⊕ # ×
1: Annotation extraction recipe Elements to extract from the annotation source	51	91: GEMINI load on dat a 90 normal tumor	⊕ ∦ ×
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		90: SnpEff eff: on data 88 normal tumor	⊕ # ×
hs_qvalue		19,382 lines, 138 commen format: vcf, database: hg	its 19
A column with the name provided here will be added to the variants table of the GEMINI database to store the annotations (-c)			
O Numbers with decimal precision       O Integer numbers       O Text (text)		display at UCSC main display with IGV local display at RViewer main	
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		##fileformat=VCFv4.2 ##FILTER= <id=pass,descriptic ##reference=/data/db/reference=</id=pass,descriptic 	on="All filters
the smallest of the (numeric) values	•	##source-varscan.py	-
	*	11	>

### Adding links to CIViC

Analyze Data Workflow Visualize * Shared Data * Help * User * 🔝 🏬		1	Using 12%
Dataset to use as the annotation source	*	History	2+⊡¢
0 0 55:01-Feb-2019-CIVic.bed	8	search datasets	00
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 option to make use of any annotation that overlaps with the position of a database variant. This setting is ignored for annotation sources in BED	ł	Tumor Normal pair somatic pipeline TEST 58 shown, 39 deleted, hide hid 30.42 GB	den
format, for which matching is always based on overlapping positions only. (region-only)		94: GEMINI annotate o	@ / x _
Specific values extracted from matching records in the annotation source (extract)		n data 53 and data 93	
(-a)	_	93: GEMINI annotate o n data 56 and data 92	⊛ # ×
1: Annotation extraction recipe  Elements to extract from the annotation source	- 1	92: GEMINI annotate o n data 88 and data 91 normal tumor	⊕ # ×
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)		91: GEMINI load on dat a 90 normal tumor	⊕ # ×
Database column name to use for recording annotations	_	90: SnpEff eff: on data	⊛ # ×
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?		88 normal tumor 19,382 lines, 138 comments	
<ul> <li>○ Numbers with decimal precision</li> <li>○ Integer numbers</li> <li>⊘ Text (text)</li> </ul>		format: vcr, database: hg 19 ③ & O 있 네 ? display at UCSC main	••
Your selection will determine the data type used to store the new annotations in the database. (-t)		display with IGV local display at RViewer main	
If multiple annotations are found for the same variant, store		1.Chron	
a comma-separated list of non-redundant (text) values	•	##T1Leformat=VCFv4.2	)

### Adding further annotations from Cancer Genome Interpreter (CGI)

Analyze Data 🛛 Workflow Visualize 🍷 Shared Data 🍷 Help 🍷 User 🍷 📻		Using 12%
GEMINI annotate the variants in an existing GEMINI database with additional information (Galaxy Version 0.20.1+galaxy2)	☆ Favorite & Versions ▼ Options	History & + □ ↓
GEMINI database         D       D         95: GEMINI annotate on data 55 and data 94         Only files with version 0.20.1 are accepted.	• 1	Tumor Normal pair somatic pipeline TEST 59 shown. 39 deleted. hide hidden 30.42 GB
D       D       54: cgi_variant_positions.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)	- 1 0	95: GEMINI annotate o n data 55 and data 94 94: GEMINI annotate o
Yes The default is to consider VCF-formatted annotations only if a variant in the GEMINI database and a record in the annotation source desc position in the genome. You can disable this option to make use of any annotation that overlaps with the position of a database variant. T format, for which matching is always based on overlapping positions only. (region-only)	ribe the exact same nucleotide change at the same 'his setting is ignored for annotation sources in BED	n data 53 and data 93 93: GEMINI annotate o n data 56 and data 92
Type of information to add to the database variants Binary indicator (1=found, 0=not found) of whether the variant had any match in the annotation source (boolean)	•	92: GEMINI annotate o 🕢 🖉 🗙 n data 88 and data 91 normal tumor
(-a) Database column name to use for recording annotations		91: GEMINI load on dat 🕢 🥒 🗙 a 90 normal tumor
in_cgidb A column with the name provided here will be added to the variants table of the GEMINI database to store the annotations (-c) Email notification		90: SnpEff eff: on data 88  normal tumor  19,382 lines, 138 comments
No Send an email notification when the job completes.		format: vcf, database: hg19 ③ ⑦ ① ② 世 ? ③ ● display at UCSC main display with IGV local display. at PC/iswer main
What it does		• II >

## b. Reporting selected subsets of variants

### Querying the GEMINI database for somatic variants

Analyze Data Workflow Visualize	Help -	User 🕶 🚺	-	
GEMINI query Querying the GEMINI database (Galaxy Version 0.20.1+galaxy1)	Favorite	le Version	ns 🗸 🕶 O	ptions
EMINI database				
🖸 🕼 🗅 96: GEMINI annotate CGI infos on data	4 and data 9	5	•	<b>t</b> 🕞
Only files with version 0.20.1 are accepted.				
uild GEMINI query using	1			
Basic variant query constructor				-
Genotype filter expression				
1: Genotype filter expression				<b></b>
Restrictions to apply to genotype values				
(gt-filter)				1
+ Insert Genotype filter expression				
Sample filter expression				
+ Insert Sample filter expression				
Region Filter				
+ Insert Region Filter				
Filter variant sites by their position in the genome. If multip fall in ONE of the regions are reported.	e Region Filte	ers are spe	cified, all v	ariants that
Additional constraints expressed in SQL syntax				
somatic_status = 2				
Constraints defined here will become the WHERE clause of database. E.g. alt='G' or impact severity = 'HIGH'.	ne SQL quer	y issued to	the GEMII	NI

Analyze Data Workflow Visualize   Shared Data Help  User		Using 12%
utput format options	History	2+□\$
/pe of report to generate	search dat	tasets 00
tabular (GEMINI default)	• Tumor No	rmal pair
Add a neader or column names to the output	somatic pi	peline TEST
Ves Ves	61 shown, 39 d	deleted, hide hidden
(beader)	31.3 GB	
Set of columns to include in the variant report table		
Custom (report user-specified columns)		NI query on 💿 🖋 🗙
Choose columns to include in the report	normal tum	or
E Select/Unselect all	This job is wa	aiting to run
□ gene ☑ chrom	0 2 ?	•
☑ start	96: GEMINI	annotate CG 💿 🖋 🗙
end Filosof	l infos on da	ita 54 and d
☑ alt	ata 95	
impact	95: GEMINI	annotate Cl 💿 🖋 🗙
impact_severity	ViC data on	data 55 and
alternative allele frequency (max_aat_all)	data 94	or
(columns)		_
Additional columns (comma-separated)	94: GEMINI	annotate Ca 💿 🖋 X
gene, aa_change, rs_ids, hs_qvalue, cosmic_ids	a 53 and dat	ta 93
Column must be specified by the exact name they have in the GEMINI database, e.g.	is exonic or	or
num_hom_alt, but, for genotype columns, GEMINI wildcard syntax is supported. The	order of O2: GEMINI	annatata dh (A) A V
columns in the list is maintained in the output.	SNP infos or	n data 56 an
Request drug-gene interaction info from DGldb	d data 92	
No No	normal tum	or
(dgidb)	92: GEMINI	annotate Va 💿 🖋 🗙
Sort the output by the following column(s)	rScan Soma	tic infos on
	data 88 and	data 91

### **GEMINI SQL-based output formatting**

💶 Galaxy Europe	Analyze Data Workflow Visualize 🕶 Shared Data 🍷 Help 👻 User 🍷 📻 🇮	
Tools රු	database. E.g. alt='G' or impact_severity = 'HIGH'.	type,
Comini guang	Output format options	
Gemini query	Type of report to generate	gt_alt_treqs.NORMAL,
1 Upload Data	tabular (GEMINI default)	<pre>itnull(nullit(round(max_aat_all,2),-1.0)</pre>
	Add a header of column names to the output	AS MAF,
🕅 Hide Sections	Yes	jene,
Gemini	(header)	Impact_S0,
GEMINI query Querying the GEMINI	Set of columns to include in the variant report table	dd_Change,
database	Custom (report user-specified columns)	ITHUII(round(cadd_scaled,2), . ) AS
GEMINI set_somatic Tag somatic mutations in a GEMINI database	Choose columns to include in the report	round(gerp bp score,2) AS gerp bp,
GEMINI fusions Identify somatic		ifnull(round(gerp element pval,2),'.') A
fusion genes from a GEMINI database GEMINI amend Amend an already loaded GEMINI database. GEMINI load Loading a VCF file into GEMINI	□ gene ☑ chrom ☑ start □ end ☑ ref	<pre>gerp_element_pval, ifnull(round(hs_qvalue,2), '.') AS hs_qvalue, in omim.</pre>
GEMINI annotate the variants in an existing GEMINI database with additional information	Impact     Impact	<pre>ifnull(clinvar_sig,'.') AS clinvar_sig, ifnull(clinvar_disease_name,'.') AS</pre>
GEMINI database info Retrieve information about tables, columns and annotation data stored in a GFMINI database	(columns) Additional columns (comma-separated)	<pre>clinvar_disease_name, ifnull(rs_ids,'.') AS dbsnp_ids, ns_ss</pre>
GEMINI stats Compute useful variant statistics	type, gt_alt_freqs.TUMOR, gt_alt_freqs.NORMAL, ifnull(nullif(round(max_aaf_all,2),-1.0),0)	<sup>()</sup> AS MAI ifnull(cosmic_ids,'.') AS cosmic_ids,
GEMINI actionable_mutations Retrieve genes with actionable	num_hom_alt, but, for genotype columns, GEMINI wildcard syntax is supported. The order columns in the list is maintained in the output.	der of overlapping_civic_url, '.') AS
DGIdb	Request drug-gene interaction info from DGldb	in_cgidb
•	No No	

### c. Generating reports of genes affected by variants

#### Turning query results into gene-centered reports

Analyze Data Workflow Visualize 🔹 Shared Data 🔹 Help 🔭 🔛 🏭	Using 12%
	History 😅 🕂 🖽
GEMINI query Querying the GEMINI database (Galaxy Version 0.20.1+galaxy1)     ☆ Favorite     & Versions <ul> <li>Options</li> </ul>	search datasets
EMINI database         Image: Description of the second s	Tumor Normal pair somatic pipeline TEST 64 shown, 39 deleted, hide hidden 31.3 GB
ild GEMINI query using	
Advanced query constructor	▼ 100: GEMINI query on d ④ A × ata 96
The query to be issued to the database 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	99: GEMINI query on da 💿 🖋 🗙 ta 96
v.somatic_status = 2 AND v.somatic_p <= 0.05 AND v.filter IS NULL GROUP BY g.gene	98: GEMINI database inf 🔹 🖋 🗙 o on data 96
Formulate your query using SQL syntax. (-q) Genotype filter expression	97: GEMINI query on da 💿 🖋 🗙 ta 96
1: Genotype filter expression	96: GEMINI annotate CG 💿 🖋 🗙
Restrictions to apply to genotype values	ata 95
gt_alt_freqs.NORMAL <= 0.05 AND gt_alt_freqs.TUMOR >= 0.10	95: GEMINI annotate CI ( ) * × ViC data on data 55 and data 94 normal tumor
(gt-filter)	94: GEMINI annotate Ca 💿 🖋 🗙
+ Insert Genotype filter expression	ncer Hotspots V2 on dat a 53 and data 93
Sample filter expression	93: GEMINI annotate dh
+ Insert Sample filter expression	SNP infos on data 56 an d data 92
Output format options	normal tumor

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

GROUP BY g.gene

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>

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

### Adding UniProt cancer genes information

📮 Galaxy Europe	Analyze Data Workflow Visualize - Shared Data - Help - User - 🎓 🏢	Using 12%
Tools ☆	Lain two files (Calaus Version 1.1.2)	Ĥistory 😂 + ⊡ 🌣
join two files	Cont two mes (Galaxy Version 1.1.2) ☆ Favorite & Versions ▼ Options	search datasets 🛛 🕄 🕄
1 Upload Data	1st file           107: GEMINI query on data 96	Tumor Normal pair somatic pipeline TEST
Show Sections	Column to use from 1st file	37 shown, 45 deleted, 29 hidden
VCF-VCFintersect: Intersect two VCF datasets	Column: 1	31.3 GB 🗹 📎 🗩
Sub-sample sequences files e.g. to reduce coverage	2nd File	🕚 108: Join on data 57 💿 🖋 🗙 and data 107
idpQuery Creates text reports from idpDB files.	Column to use from 2nd file	This job is waiting to run 9 것 ? >>>>
Join two files	Column: 1	
fastq-join - Joins two paired-end reads on the overlapping ends	Output lines appearing in	107: GEMINI query on d 💿 🖋 🗙 ata 96
Join two files on column allowing a small difference	Both 1st & 2nd file, plus unpairable lines from 1st file. (-a 1)	43 lines format: <b>tabular</b> , database: <b>hg19</b>
Join the intervals of two datasets side- by-side	First fine is a header line	800211? >>
Column Join on Collections	Use whist line contains column headers. It will not be sorted.	gene chrom synonym hgnc_i
Column Join on Collections	enore care	A2ML1 chr12 CPAMD9,FLJ25179 23336
QCMerger Merges two qcml files together.	No Soft and Join key column values regardless of upper/lower case letters	ANKDD1B chr5 None 32525 APC chr5 DP3,PPP1R46,DP2,DP2.5 583 ARHGAP9 chr12 18C MGC1295 14138
FuzzyDiff Compares two files, tolerating numeric differences.	Value to put in unpaired (empty) fields	<
seqtk_mergefa merge two FASTA/Q files	0 Eman polification	100: GEMINI query on d (*) / × ata 96
Join +/- lons Join positive and negative ionization-mode W4M	No No	99: GEMINI query on da 💿 🖋 🗙
datasats for the same samples	Send an email notification when the job completes.	

### Adding CGI biomarkers information

📮 Galaxy Europe	Analyze Data Workflow Visualize 🕶 Shared Data 🍷 Help 🍷 User 👻 📑	Using 12%
Tools ☆	Join two files (Galaxy Version 1.1.2)	History 🕃 🕂 🗆 🌣
join two files 😢		search datasets
🍰 Upload Data	1st file           D         D           102: Join on data 57 and data 100	Tumor Normal pair somatic pipeline TEST
Show Sections	Column to use from 1st file	66 shown, 39 deleted, hide hidden
VCF-VCFintersect: Intersect two VCF	Column: 1	31.3 GB
Sub-sample sequences files e.g. to reduce coverage	2nd File	102: Join on data 57 and 💿 🖋 🗙 data 100
idpQuery Creates text reports from idpDB files.		70 lines
Join two files	Column: 1	format: tabular, database: hg19
fastq-join - Joins two paired-end reads on the overlapping ends	Output lines appearing in	
Join two files on column allowing a small difference	Both 1st & 2nd file, plus unpairable lines from 1st file. (-a 1)	gene chrom start ref alt type MAF i A2ML1 chr12 8989045 A C snp 0 i
Join the intervals of two datasets side- by-side	Electine is a header line	ANKDD1B chr5 74930649 G A snp 0 1 APC chr5 112175422 C T snp 0 s
Column Join on Collections	de un first line contains column headers. It will not be sorted.	4
Column Join on Collections	Ignoraciase	101: Join on data 57 and () A X
QCMerger Merges two qcml files together.	Soft and Join key column values renardless of unner/Jower case letters	data 100 normal tumor
FuzzyDiff Compares two files, tolerating numeric differences.	Value to put in unpaired (empty) fields	70 lines format: <b>tabular</b> , database: <b>hg19</b>
<b>seqtk_mergefa</b> merge two FASTA/Q files		₿₽₿₽₩? ♥♥
Join +/- lons Join positive and negative ionization-mode W4M	No No	1 2 3 4 5 6 7 chrom start ref alt type MAF gene chr12 113335665 6 A snn 0.0 8PH3A
datasets for the same samples	Send an email notification when the job completes.	chr12 117582871 C CT indel 0.29 FBX021
Multi-Join (combine multiple files)	✓ Execute	chr12 123286211 G A snp 0.0 CCDC62 chr12 126068542 A G snp 0 TMEM132
<		- III

### Adding gene information from CIViC

📮 Galaxy Europe	Analyze Data Workflow Visualize 🕶 Shared Data 🍷 Help 👻 User 👻 📻 🇮		Using 12%
Tools ☆		er	
join two files	IFOpe       Analyze Data       Workflow       Visualize *       Shared Data *       Help *       User *       Image: Control of the state in the		
📩 Upload Data	1st file	Tumor Normal pair somatic pipeline TEST	
Show Sections	Column to use from 1st file	39 shown, 45 deleted, 29 hidder	n
VCF-VCFintersect: Intersect two VCF	Column: 1	31.3 GB	
Sub-sample sequences files e.g. to reduce coverage	2nd File	110: Join on data 59 and data 109	● # ×
idpQuery Creates text reports from idpDB files.	Column to use from 2nd file	This job is waiting to run の	۰
Join two files	Column: 3		
fastq-join - Joins two paired-end reads on the overlapping ends	Output lines appearing in	109: Join on data 58 and data 108	● / ×
Join two files on column allowing a small difference	Both 1st & 2nd file, plus unpairable lines from 1st file. (-a 1)	108: Join on data 57 and data 107	● / ×
Join the intervals of two datasets side- by-side	Yes	107: GEMINI query on d	● # ×
Column Join on Collections	Use if inst line contains column headers. It will not be sorted.	ata 96	
Column Join on Collections	gnore case	100: GEMINI query on d	● / ×
QCMerger Merges two qcml files together.	No Soft and Join key column values regardless of upper/lower case letters	normal tumor	
FuzzyDiff Compares two files, tolerating numeric differences.	Value to put in unpaired (empty) fields	99: GEMINI query on da ta 96	● # ×
seqtk_mergefa merge two FASTA/Q files		98' GEMINI database inf	• A X
Join +/- Ions Join positive and negative ionization-mode W4M	No     Send an amail polification when the job completes	o on data 96	~ , A

### Rearrange to get a fully annotated gene report

🗧 Galaxy Europe	Analyze Data Workflow Visualize 🕶 Shared Data 🍷 Help 👻 User 🍸 💼	Analyze Data Workflow Visualize - Shared Data - Help - User - 🞓 🇮
Tools		6: Specify the first few columns by name
Column arrange	Column arrange by header name (Galaxy Version     A Favorite     A Versions      Options	column
🛓 Upload Data	file to rearrange	7: Specify the first few columns by name
Show Sections	Specify the first few columns by name	column
Column arrange by header name	1: Specify the first few columns by name	IS_OG
FASTO Summary Statistics by column	column	8: Specify the first few columns by name
BLAST parser Convert 12- or 24- column BLAST output into 3-column	gene	is_TS
hcluster_sg input	2: Specify the first few columns by name	9: Specify the first few columns by name
Sort Column Order by heading Split file according to the values of a column	chrom	column
Add input name as column to an existing tabular file	3: Specify the first few columns by name	10: Specify the first few columns by name
Replace Text in a specific column	column	column
Replace column by values which are defined in a convert file	synonym	clinvar_gene_phenotype
FASTQ Trimmer by column	4: Specify the first few columns by name	11: Specify the first few columns by name
Column Join on Collections	column	column
Column Join on Collections	hgnc_id	gene_civic_url
Histogram of a numeric column	5: Specify the first few columns by name	
Map peptides to a bed file for viewing in a genome browser	column	12: Specify the first few columns by name column
Manipulate loom object Add layers, or row/column attributes to a loom file	entrez_id	description
Summary Statistics for any numerical	6: Specify the first few columns by name	+ Insert Specify the first few columns by name

### Inspecting fully annotated gene report

	gene	chrom	synonym	hgnc_id	entrez_id rvis_pct	is_OG	is_TS	in_cgi_biom clinvar_gen	gene_civic_u	description	gene_id	entrez_id	last_review	/_date	
	A2ML1	chr12	CPAMD9,FLI25179	23336	144568 98.52559566	0	) (	0 nonsyndrom							
	ANKDD1B	chr5	None	32525	728780 None	0	) (	0 None							
Г	APC	chr5	DP3,PPP1R46,DP2	2 583	324 0.902335456	0	) 1	. 1 apc-associa	https://civic		66	i 324	2017-02-09	21:58:08 UTC	
	ARHGAP9	chr12	10C,MGC1295	14130	64333 30.90351498	0	) (	0 coronary_ar	1.						
	C2CD5	chr12	CDP138,KIAA0528	3 29062	9847 None	0	) (	0 None							
	CCDC62	chr12	ERAP75,CT109,FL	J 30723	84660 82.29535268	0	) (	0 None							
	CDH12	chr5	Br-cadherin,CDH	1751	1010 29.48808681	0	) (	0 None							
	CDH18	chr5	EY-CADHERIN,CD	1757	1016 5.602736494	0	) (	0 None							
	CLEC4C	chr12	DLEC,CLECSF7,CD	13258	170482 86.47676339	0	) (	0 None							
	CLEC6A	chr12	dectin-2,CLECSF1	14556	93978 89.95635763	0	) (	0 None							
	COX7C	chr5	None	2292	1350 62.38499646	0	) (	0 None							
	DDX51	chr12	None	20082	317781 64.96225525	0	) (	0 None							
	ELAC2	chr17	FLJ10530,HPC2	14198	60528 10.12031139	1		0 combined o							
-	ERN1	chr17	IRE1P, IRE1	3449	2081 24.63434772	0	) (	0 None							
	ESM1	chr5	None	3466	11082 56.64071715	0	) (	0 None							
	FBXO21	chr12	FBX21,KIAA0875	13592	23014 12.77423921	0	) (	0 None							
	HAPLN1	chr5	CRTL1	2380	1404 20.53550366	0	) (	0 None							
	ITGB7	chr12	None	6162	3695 4.62962963	0	) (	0 None							
	KRAS	chr12	KRAS1,KRAS2	6407	3845 42.87567823	1		1 acute_myel	https://civic	Mutations i	i 30	3845	2017-02-09	21:59:28 UTC	
	KRBA2	chr17	None	26989	124751 57.31304553	0	) (	0 None							
	LINC01019	chr5	None	27742	285577 None	0	) (	0 None		-					
	LYRM7	chr5	FLI20796,C5orf31	, 28072	90624 56.2514744	0	) (	0 mitochondri							
	METTL2A	chr17	METTL2,FLI12760	25755	339175 67.03231894	0	) (	0 None							
	MROH2B	chr5	FLI40243,DKFZp7	26857	133558 None	0	) (	0 None							
	PCDHB9	chr5	PCDH-BETA9,PCD	8694	None None	0	) (	0 None							
	PCDHGB1	chr5	PCDH-GAMMA-B1	1 8708	56104 10.92238736	0	) (	0 None							
	PCDHGB7	chr5	PCDH-GAMMA-B	7 8714	56099 48.90894079	0	) (	0 None							
	PDE3A	chr12	CGI-PDE	8778	5139 6.894314697	0	) (	0 brachydacty							
	RACK1	chr5	GNB2L1.H12.3.Gr	None	10399 46.20193442	0	) (	0 None							
	RNF213	chr17	NET57,KIAA1618,I	14539	57674 97.65274829	1	. (	0 moyamoya_							
	SLC16A5	chr17	MCT5,MCT6	10926	9121 36.22906346	0	) (	0 None							
	SMARCC2	chr12	Rsc8,CRACC2,BAF	11105	6601 2.707006369	0	) (	0 malignant_t	ι.						
	SOX5	chr12	L-SOX5,MGC3515	11201	6660 9.088228356	0	) (	0 aplasia/hyp							
	SPEF2	chr5	KPL2,FLI23577,CT	26293	79925 99.06817646	0	) (	0 None							
	SYNPO	chr5	KIAA1029	30672	11346 40.67586695	0	) (	0 None		-					
	TENM2	chr5	Ten-M2,KIAA1127	29943	57451 None	0	) (	0 None							
	TMEM132B	chr12	KIAA1906.KIAA17	29397	114795 2.677518283	0	0	0 None							
	TP53	chr17	LFS1,p53	11998	7157 35.98726115	0	) 1	1 acute_mega	https://civic	TP53 mutati	( 45	7157	2018-03-30	15:05:39 UTC	
	TTC23L	chr5	FLI25439	26355	153657 96.55579146	0	) (	0 None							
	TTC37	chr5	KIAA0372	23639	9652 58.00896438	0	) (	0 malignant_t	t.						
	USP22	chr17	KIAA1063,USP3L	12621	23326 27.41802312	0	) (	0 None							
	VCAN	chr5	CSPG2.PG-M	2464	1462 19.95753715	0	) (	0 malignant 1							