

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


Shared Data → Histories


bilille_TP2_GTN_Somatic_Variants

2.61 GB


 ? x

Dataset	Annotation
9: sorted.corrected.01-Feb-2019-CIVic.bed	
8: cgi_genes.txt	
7: 01-Feb-2019-GeneSummaries.tsv	
6: dbsnp.b147.chr5_12_17.vcf.gz	
5: 01-Feb-2019-CIVic.bed	
4: cgi_variant_positions.bed	
3: hotspots.bed	
2: Map with BWA-MEM on tumor reads	
1: Map with BWA-MEM on normal reads	

About this History 

Author
eag 

Related Histories
All published histories
Published histories by eag


Rating
Community
(0 ratings, 0.0 average) 




Tags
Community:
none



Prepare Data








💡 Tip: Adding a tag


- 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

2: Map with BWA-MEM   
on tumor reads

1.3 GB
format: **bam**, database: **hg19**




[M::mem_pestat] analyzing insert size distribution for orientation FF...
[M::mem_pestat] (25, 50, 75) percentile: (69, 104, 143)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 291)
[M::mem_pestat] mean and std.dev: (99.08, 44)

 **Edit dataset tags**








display at UCSC main test
display at Ensembl Current
display with IGV local
display in IGB View


Binary bam alignments file

2: Map with BWA-MEM   
on tumor reads

1.3 GB
format: **bam**, database: **hg19**




[M::mem_pestat] analyzing insert size distribution for orientation FF...
[M::mem_pestat] (25, 50, 75) percentile: (69, 104, 143)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 291)
[M::mem_pestat] mean and std.dev: (99.08, 44)


      



display at UCSC main test
display at Ensembl Current
display with IGV local
display in IGB View








Binary bam alignments file




2: Map with BWA-MEM   
on tumor reads

tumor 

1.3 GB
format: **bam**, database: **hg19**

[M::mem_pestat] analyzing insert size distribution for orientation FF...
[M::mem_pestat] (25, 50, 75) percentile: (69, 104, 143)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 291)
[M::mem_pestat] mean and std.dev: (99.08, 44)

#tumor   

1. Mapped reads postprocessing

1. Mapped reads postprocessing

- a. Filtering on mapped reads properties

Filtering for mapping status and quality

Galaxy France Using 18%

! From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools ☆ ☰

search tools ✕

Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION

Convert Formats

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

VCF/BCF

Nanopore

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

Fetch Alignments/Sequences

GENOMICS ANALYSIS

Filter BAM datasets on a variety of attributes (Galaxy Version 2.5.1+galaxy0) ☆ 🔗 ▼

BAM dataset(s) to filter

28: MarkDuplicates on filtered tumor BAM

26: MarkDuplicates on filtered normal BAM

19: Filtered tumor BAM

17: Filtered normal BAM

15: Map with BWA-MEM on tumor reads

14: Map with BWA-MEM on normal reads

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Condition

1: Condition

Filter

1: Filter 🗑️

Select BAM property to filter on

Mapping quality

Filter on read mapping quality (phred scale)

>=1

You can use >, <, =, and ! (not) in your expression. E.g., to select reads with mapping quality of at least 30 use ">=30"

2: Filter **"Insert Filter"**

Select BAM property to filter on

Mapped

Selected mapped reads

Yes

Checked = Mapped, Empty = NOT mapped

3: Filter

History ↺ + 🗑️ ⚙️

search datasets ? ✕

test_Somatic

53 shown, 9 deleted, 1 hidden

18.47 GB ☑️ 👤 💬

24: RmDup on data 14 👁️ ✎ ✕

19: Filtered tumor BAM 👁️ ✎ ✕

18: Filter on data 15: JS ON filter rules 👁️ ✎ ✕

17: Filtered normal BAM 👁️ ✎ ✕

16: Filter on data 14: JS ON filter rules 👁️ ✎ ✕

15: Map with BWA-MEM on tumor reads 👁️ ✎ ✕

14: Map with BWA-MEM on normal reads 👁️ ✎ ✕

11: Trimmomatic on SL GFSK-T_231336_r2_chr5_12_17.fastq.gz (R2 paired) 👁️ ✎ ✕

10: Trimmomatic on SL GFSK-T_231336_r1_chr5 👁️ ✎ ✕

In "3: Filter":

- **"Select BAM property to filter on":** isMateMapped
 - **"Select reads with mapped mate":** Yes

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

Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region (Galaxy Version 1.8+galaxy1) ☆ Favorite 🗑 Versions ▾ Options

SAM or BAM file to filter

87: Filter on data 79: Filtered BAM

Header in output

Include header

Minimum MAPQ quality score

1

(-q)

Skip alignments with any of these flag bits set

Select/Unselect all

- Read is paired
- Read is mapped in a proper pair
- The read is unmapped
- The mate is unmapped
- 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

(-F)

Mapped reads postprocessing

- b. Removing duplicate reads

Remove duplicates with MarkDuplicates

Galaxy France Workflow Visualize Shared Data Help User Using 18%

! From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools

Filter SAM

Upload Data

Show Sections

Samtools view - reformat, filter, or subsample SAM, BAM or CRAM

bcftools filter Apply fixed-threshold filters

GEMINI query Querying the GEMINI database

Kraken-filter filter classification by confidence score

msPurity.filterFragSpectra Filter fragmentations spectra associated with an XCMS feature

Snpsift Filter Filter variants using arbitrary expressions

Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region

Filter SAM on bitwise flag values

Sub-sample sequences files e.g. to reduce coverage

VCFfilter: filter VCF data in a variety of attributes

GEMINI lof_sieve Filter LoF variants by transcript position and type

Filter genes with rare and low

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.18.2.3)

Select SAM/BAM dataset or dataset collection

28: MarkDuplicates on filtered tumor BAM
26: MarkDuplicates on filtered normal BAM
19: Filtered tumor BAM
17: Filtered normal BAM
15: Map with BWA-MEM on tumor reads
14: Map with BWA-MEM on normal reads

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

If empty, upload or import a SAM/BAM dataset

Comment

+ Insert Comment

You can provide multiple comments

If true do not write duplicates to the output file instead of writing them with appropriate flags set

Yes

REMOVE_DUPLICATES; default=False

Assume the input file is already sorted

Yes

ASSUME_SORTED; default=True

The scoring strategy for choosing the non-duplicate among candidates

SUM_OF_BASE_QUALITIES

Duplicate_Scoring_Strategy; default=SUM_OF_BASE_QUALITIES

Regular expression that can be used in unusual situations to parse non-standard read names in the incoming SAM/BAM dataset

READ_NAME_REGEX; Read names are parsed to extract three variables: tile/region, x coordinate and y coordinate. These values are used to estimate the rate of optical duplication in order to give a more accurate estimated library size. See help below for more info; default="" (uses : separation)

History

search datasets

test_Somatic

53 shown, 9 deleted, 1 hidden

18.47 GB

19: Filtered tumor BAM
tumor

18: Filter on data 15: JS
ON filter rules

17: Filtered normal BA
M
normal

16: Filter on data 14: JS
ON filter rules

15: Map with BWA-ME
M on tumor reads
tumor

14: Map with BWA-ME
M on normal reads
normal

11: Trimmomatic on SL
GFSK-T_231336_r2_chr5
_12_17.fastq.gz (R2 paired)

10: Trimmomatic on SL
GFSK-T_231336_r1_chr5
_12_17.fastq.gz (R1 paired)

Mapped reads postprocessing

- c. Left-align reads around indels

Left-align with BamLeftAlign

Galaxy France Workflow Visualize Shared Data Help User Using 18%

From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools

Filter SAM

Upload Data

Show Sections

Samtools view - reformat, filter, or subsample SAM, BAM or CRAM

bcftools filter Apply fixed-threshold filters

GEMINI query Querying the GEMINI database

Kraken-filter filter classification by confidence score

msPurity.filterFragSpectra Filter fragmentations spectra associated with an XCMS feature

SnpSift Filter Filter variants using arbitrary expressions

Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region

Filter SAM on bitwise flag values

Sub-sample sequences files e.g. to reduce coverage

VCFfilter: filter VCF data in a variety of attributes

GEMINI lof_sieve Filter LoF variants by transcript position and type

Filter genes with rare and low

BamLeftAlign indels in BAM datasets (Galaxy Version 1.3.6)

Choose the source for the reference genome

Locally cached

Select alignment file in BAM format

- 28: MarkDuplicates on filtered tumor BAM
- 26: MarkDuplicates on filtered normal BAM
- 19: Filtered tumor BAM
- 17: Filtered normal BAM
- 15: Map with BWA-MEM on tumor reads
- 14: Map with BWA-MEM on normal reads

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Using reference genome

Human (Homo sapiens): hg19

(--fasta-reference)

Maximum number of iterations

5

Iterate the left-realignment no more than this many times (--max-iterations)

Email notification

Send an email notification when the job completes.

Execute

When calling indels, it is important to homogenize the positional distribution of insertions and deletions in the input by using left realignment. Left realignment will place all indels in homopolymer and microsatellite repeats at the same position, provided that doing so does not introduce mismatches between the read and reference other than the indel. This method is computationally inexpensive and handles the most common classes of alignment inconsistency.

This is leftalign utility from FreeBayes package.

History

search datasets

test_Somatic

52 shown, 10 deleted, 1 hidden

18.47 GB

- 28: MarkDuplicates on filtered tumor BAM
- 27: MarkDuplicates on data 19: MarkDuplicate metrics
- 26: MarkDuplicates on filtered normal BAM
- 25: MarkDuplicates on data 17: MarkDuplicate metrics
- 19: Filtered tumor BAM **tumor**
- 18: Filter on data 15: JS ON filter rules
- 17: Filtered normal BAM **normal**
- 16: Filter on data 14: JS ON filter rules
- 15: Map with BWA-MEM

WHY LEFT ALIGN???

Mapped reads postprocessing

- d. Recalibrate read mapping qualities

Recalibrate read quality scores with CalMD

Galaxy France

Workflow Visualize Shared Data Help User Using 18%

From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools

calmd

Upload Data

Show Sections

Samtools calmd recalculate MD/NM tags

WORKFLOWS

All workflows

Samtools calmd recalculate MD/NM tags (Galaxy Version 2.0.3)

BAM file to recalculate

30: BamLeftAlign on tumor data (alignments)
29: BamLeftAlign on normal data (alignments)
28: MarkDuplicates on filtered tumor BAM
26: MarkDuplicates on filtered normal BAM
19: Filtered tumor BAM
17: Filtered normal BAM
15: Map with BWA-MEM on tumor reads

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Use a reference sequence

Use a built-in genome

Reference

Human (Homo sapiens): hg19

Do you also want BAQ (Base Alignment Quality) scores to be calculated?

No

(-r)

Additional options

Advanced options

Change identical bases to '='

No

Replace bases in read sequences that match the reference base at that position with an equal sign (-e)

Coefficient to cap mapping quality of poorly mapped reads

50

Higher values for this setting mean a stronger downgrade of the mapping quality of reads with excessive mismatches (50: recommended setting for reads aligned with BWA, 0: do not downgrade mapping qualities) (-C)

Email notification

History

search datasets

test_Somatic

52 shown, 10 deleted, 1 hidden

18.47 GB

30: BamLeftAlign on tumor data (alignments) tumor

29: BamLeftAlign on normal data (alignments) normal

28: MarkDuplicates on filtered tumor BAM

27: MarkDuplicates on data 19: MarkDuplicate metrics

26: MarkDuplicates on filtered normal BAM

25: MarkDuplicates on data 17: MarkDuplicate metrics

19: Filtered tumor BAM tumor

18: Filter on data 15: JS ON filter rules

Mapped reads postprocessing

e. Refilter reads based on mapping quality

Eliminating reads with undefined mapping quality

Galaxy France Workflow Visualize Shared Data Help User Using 18%

! From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools calmd Upload Data Show Sections Samtools calmd recalculate MD/NM tags WORKFLOWS All workflows

Filter BAM datasets on a variety of attributes (Galaxy Version 2.5.1+galaxy0)

BAM dataset(s) to filter

- 36: Samtools calmd on tumor data
- 35: Samtools calmd on normal data
- 30: BamLeftAlign on tumor data (alignments)
- 29: BamLeftAlign on normal data (alignments)
- 28: MarkDuplicates on filtered tumor BAM
- 26: MarkDuplicates on filtered normal BAM
- 10: Filtered tumor BAM

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Condition

1: Condition

Filter

1: Filter

Select BAM property to filter on

Mapping quality

Filter on read mapping quality (phred scale)

<=254

You can use >, <, =, and ! (not) in your expression. E.g., to select reads with mapping quality of at least 30 use ">=30"

+ Insert Filter

+ Insert Condition

Would you like to set rules?

No

Allows complex logical constructs. See Example 4 below.

Email notification

History search datasets test_Somatic 52 shown, 10 deleted, 1 hidden 18.47 GB

- 36: Samtools calmd on tumor data tumor
- 35: Samtools calmd on normal data normal
- 34: dbsnp.b147.chr5_12_17.vcf.gz
- 33: 01-Feb-2019-CIVic.bed
- 32: cgi_variant_positions.bed
- 31: hotspots.bed
- 30: BamLeftAlign on tumor data (alignments) tumor
- 29: BamLeftAlign on normal data (alignments) normal
- 28: MarkDuplicates on f

2. Variant calling and classification

Variant calling with VarScan somatic

Galaxy France

Workflow Visualize Shared Data Help User

Using 18%

From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools

calmd

Upload Data

Show Sections

Samtools calmd recalculate MD/NM tags

WORKFLOWS

All workflows

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?

Use a built-in genome

reference genome

Human (Homo sapiens): hg19

The fasta reference genome that variants should be called against.

aligned reads from normal sample

39: Filtered calmd normal BAM

aligned reads from tumor sample

41: Filtered calmd tumor BAM

Estimated purity (non-tumor content) of normal sample

1

(--normal-purity)

Estimated purity (tumor content) of tumor sample

0.5

(--tumor-purity)

Generate separate output datasets for SNP and indel calls?

No

Settings for Variant Calling

Customize settings

Read selection

Minimum base quality

History

search datasets

test_Somatic

52 shown, 10 deleted, 1 hidden

18.47 GB

41: Filtered calmd tumor BAM
tumor

40: Filter on data 36: JS ON filter rules

39: Filtered calmd normal BAM
normal

38: Filter on data 35: JS ON filter rules

36: Samtools calmd on tumor data
tumor

35: Samtools calmd on normal data
normal

34: dbsnp.b147.chr5_12_17.vcf.gz

33: 01-Feb-2019-CIVic.bed

Variant calling with VarScan somatic

Galaxy France Workflow Visualize Shared Data Help User Using 18%

! From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools

calmd

Upload Data

Show Sections

Samtools calmd recalculate MD/NM tags

WORKFLOWS

All workflows

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?

Use a built-in genome

reference genome

Human (Homo sapiens): hg19

The fasta reference genome that variants should be called against.

aligned reads from normal sample

39: Filtered calmd normal BAM

aligned reads from tumor sample

41: Filtered calmd tumor BAM

Estimated purity (non-tumor content) of normal sample

1

(--normal-purity)

Estimated purity (tumor content) of tumor sample

0.5

(--tumor-purity)

Generate separate output datasets for SNP and indel calls?

No

Settings for Variant Calling

Customize settings

Read selection

Minimum base quality

28

The minimum base quality (default: 13) at a given position required to use a read for calling variants at that site (samtools mpileup -Q)

Minimum mapping quality

1

The minimum mapping quality (default: 0) required for a read to be considered in variant calling (samtools mpileup -q)

History

search datasets

test_Somatic

52 shown, 10 deleted, 1 hidden

18.47 GB

41: Filtered calmd tumor BAM

tumor

40: Filter on data 36: JS ON filter rules

39: Filtered calmd normal BAM

normal

38: Filter on data 35: JS ON filter rules

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

The screenshot shows the Galaxy France interface for the SnpEff tool. The tool is configured to annotate variants. Several settings are highlighted with red boxes:

- Sequence changes (SNPs, MNPs, InDels):** Set to "42: VarScan somatic on data 41 and data 39".
- Input format:** Set to "VCF".
- Output format:** Set to "VCF (only if input is VCF)".
- Genome source:** Set to "Download on demand".
- SnpEff Genome Version Name (e.g. GRCh38.86):** Set to "hg19".
- Produce Summary Stats:** Set to "No" with the option "(-noStats)" visible below.

The right sidebar shows the job history for "test_Somatic", listing several jobs with their respective filter rules and annotations (e.g., "tumor", "normal").

- a. Adding annotations to the called variants
 - a.2. Adding genetic and clinical evidence-based annotations

Creating a GEMINI database from a variants dataset

From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools

gemini

Upload Data

Show Sections

GEMINI load Loading a VCF file into GEMINI (Galaxy Version 0.20.1+galaxy2)

GEMINI set_somatic Tag somatic mutations in a GEMINI database

GEMINI gene_wise Discover per-gene variant patterns across families

GEMINI amend Amend an already loaded GEMINI database.

GEMINI fusions Identify somatic fusion genes from a GEMINI database

GEMINI annotate the variants in an existing GEMINI database with additional information

GEMINI stats Compute useful variant statistics

GEMINI query Querying the GEMINI database

GEMINI burden perform sample-wise gene-level burden calculations

GEMINI actionable_mutations Retrieve genes with actionable somatic mutations via COSMIC and DGIdb

GEMINI interactions Find genes among variants that are interacting partners

GEMINI database info Retrieve information about tables, columns and annotation data stored in a GEMINI database

GEMINI roh Identifying runs of homozygosity

GEMINI inheritance pattern based

GEMINI load Loading a VCF file into GEMINI (Galaxy Version 0.20.1+galaxy2)

VCF dataset to be loaded in the GEMINI database

47: SnpEff eff: on data 46 and data 42

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

The variants in this input are

annotated with snpEff

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 without annotations. (-t)

This input comes with genotype calls for its samples

Yes

This is usually the case, but some published datasets, like some 1000G VCFs, are missing genotype information. (--no-genotypes)

Choose a gemini annotation source

GEMINI annotations w/ GERP & CADD (2022-03-23 snapshot)

Sample and family information in PED format

Nothing selected

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 details). If you choose to not provide sample information now, but later find that you need it for your analysis, you can also add it to an existing GEMINI database by using the GEMINI amend tool. (-p)

Load the following optional content into the database

Select/Unselect all

- GERP scores
- CADD scores (non-commercial use only; see licensing note below)
- Gene tables
- Sample genotypes
- Genotype likelihoods (sample PLs)
- only variants that passed all filters
- variant INFO field

The preselected defaults should be ok for most use cases (feel free to enable CADD scores for non-commercial use). If you are not interested in certain annotations, you can speed up database creation and decrease the resulting database size slightly by not loading them into the database. Note: GERP and CADD scores are optional parts of the annotation source and can only be loaded if available.

Email notification

Send an email notification when the job completes.

Execute

History

search datasets

test_Somatic

38 shown, 9 deleted

17.14 GB

47: SnpEff eff: on data 46 and data 42

46: SnpEff download: SnpEff4.3 hg19

45: 01-Feb-2019-GeneSummaries.tsv

44: cgi_genes.txt

43: Uniprot_Cancer_Genes.13Feb2019.txt

42: VarScan somatic on data 41 and data 39

41: Filtered calmd tumor BAM

40: Filter on data 36: JS ON filter rules

39: Filtered calmd normal BAM


38: Filter on data 35: JS ON filter rules

36: Samtools calmd on tumor data

35: Samtools calmd on normal data

34: dbsnp.b147.chr5_12_17.vcf.gz

Making variant call statistics accessible

 **GEMINI annotate** the variants in an existing GEMINI database with additional information (Galaxy Version 0.20.1+galaxy2) ☆ ▾

GEMINI database

   48: GEMINI load on data 47 ▾



Only files with version 0.20.1 are accepted.

Dataset to use as the annotation source

   42: VarScan somatic on data 41 and data 39 ▾



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 format, for which matching is always based on overlapping positions only. (--region-only)

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

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

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?

- 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 ▾

Making variant call statistics accessible

2: Annotation extraction recipe

Elements to extract from the annotation source

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)

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

Elements to extract from the annotation source

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?


- 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

Adding further annotations from dbSNP

 **GEMINI annotate** the variants in an existing GEMINI database with additional information (Galaxy Version 0.20.1+galaxy2) ☆ ▾

GEMINI database

⬇️ ⬆️ ⬇️

Only files with version 0.20.1 are accepted.

Dataset to use as the annotation source

⬇️ ⬆️ ⬇️

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 format, for which matching is always based on overlapping positions only. (--region-only)

Type of information to add to the database variants

▾

(-a)

Annotation extraction recipe

1: Annotation extraction recipe

Elements to extract from the annotation source

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

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?

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 ...

▾

Note: If indicated (in parentheses) an option is only applicable to annotations of a specific type. (-o)

Adding further annotations from Cancer Hotspots v2

 **GEMINI annotate** the variants in an existing GEMINI database with additional information (Galaxy Version 0.20.1+galaxy2) ☆ ▾

GEMINI database

   50: GEMINI annotate on data 34 and data 49 ↓  

Only files with version 0.20.1 are accepted.

Dataset to use as the annotation source

   31: hotspots.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 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)

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

5

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

hs_qvalue

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?

- 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 smallest of the (numeric) 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

 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



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 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)

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?


- 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 ...

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?

Galaxy France

Workflow Visualize Shared Data Help User Using 18%

From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance

Tools

search tools

Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION

Convert Formats

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

VCF/BCF

Nanopore

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

Fetch Alignments/Sequences

GENOMICS ANALYSIS

Tool Parameters

Input Parameter	Value
GEMINI database	51 GEMINI annotate on data 31 and data 50
Dataset to use as the annotation source	33 01-Feb-2019-CIVic.bed
Strict variant-identity matching of database and annotation records (VCF format only)	True
Type of information to add to the database variants	extract
Elements to extract from the annotation source	4
Database column name to use for recording annotations	overlapping_civic_url
What type of data are you trying to extract?	Text (text)
If multiple annotations are found for the same variant, store ...	a comma-separated list of non-redundant (text) values

Job Outputs

Tool Outputs	Dataset
GEMINI annotate	52 GEMINI annotate on data 33 and data 51 (Hidden)

Job Information

Galaxy Tool ID: toolshed.g2.bx.psu.edu/repos/iuc/gemini_annotate/gemini_annotate/0.20.1+galaxy2

Command Line: empty

Tool Standard Output: empty

Tool Standard Error: [E::hts_idx_push] Invalid record on sequence #12: end 113992971 < begin 114036521
tbx_index_build failed: tabixed.bed.gz

Tool Exit Code: 1

History

search datasets

test_Somatic

63 shown, hide deleted

18.47 GB

52: GEMINI annotate on data 33 and data 51

error

An error occurred with this dataset:
[E::hts_idx_push] Invalid record on seq
tbx_index_build failed: tabixed.bed.gz

WARNING:galaxy.datatypes.binary:<ga

51: GEMINI annotate on data 31 and data 50

226.4 MB

format: gemini.sqlite, database: hg19

finished updating 4 variants

Gemini SQLite Database, version 0.20.1

50: GEMINI annotate on data 34 and data 49

49: GEMINI annotate on

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

Only files with version 0.20.1 are accepted.

Dataset to use as the annotation source

53: sorted.corrected.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 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)

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?

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 ...






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)

Adding further annotations from Cancer Genome Interpreter (CGI)






 **GEMINI annotate** the variants in an existing GEMINI database with additional information (Galaxy Version 0.20.1+galaxy2) ☆ ▾

GEMINI database

   54: GEMINI annotate on data 53 and data 51 ▾  

Only files with version 0.20.1 are accepted.

Dataset to use as the annotation source

   32: 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)

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 format, for which matching is always based on overlapping positions only. (--region-only)

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) ▾

(-a)

Database column name to use for recording annotations


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


No


Send an email notification when the job completes.




What it does

Given an existing GEMINI database and an annotation source in BED or VCF format, the annotate tool will, for each variant in the variants table of the database, screen for overlapping regions defined in the annotation source and update one or more new columns of the variant record in the database based on the result and the annotation found.

Citations: 

- Paila, U., Chapman, B. A., Kirchner, R., & Quinlan, A. R. (2013). GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Computational Biology*, 9(7), e1003153. <https://doi.org/10.1371/journal.pcbi.1003153> 

Requirements: 

- gemini (Version 0.20.1)

b. Reporting selected subsets of variants

Querying the GEMINI database for somatic variants

GEMINI query Querying the GEMINI database (Galaxy Version 0.20.1+galaxy1) Favorite Versions Options

GEMINI database

96: GEMINI annotate CGI infos on data 54 and data 95

Only files with version 0.20.1 are accepted.

Build GEMINI query using

Basic variant query constructor

Genotype filter expression

1: Genotype filter expression

Restrictions to apply to genotype values

gt_alt_freqs.NORMAL <= 0.05 AND gt_alt_freqs.TUMOR >= 0.10

(--gt-filter)

+ 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 multiple Region Filters are specified, all variants that fall in ONE of the regions are reported.

Additional constraints expressed in SQL syntax

somatic_status = 2

Constraints defined here will become the WHERE clause of the SQL query issued to the GEMINI database. E.g. alt='G' or impact_severity = 'HIGH'.

Output format options

Type of report to generate

tabular (GEMINI default)

Add a header of column names to the output

Yes

(--header)

Set of columns to include in the variant report table

Custom (report user-specified columns)

Choose columns to include in the report

Select/Unselect all

- gene
- chrom
- start
- end
- ref
- alt
- impact
- impact_severity
- alternative allele frequency (max_aaf_all)

(--columns)

Additional columns (comma-separated)

gene, aa_change, rs_ids, hs_qvalue, cosmic_ids

Column must be specified by the exact name they have in the GEMINI database, e.g., is_exonic or num_hom_alt. but, for genotype columns, GEMINI wildcard syntax is supported. The order of columns in the list is maintained in the output.

Request drug-gene interaction info from DGIdb

No

(--dgidb)

Sort the output by the following column(s)

GEMINI SQL-based output formatting

Tools ☆

Gemini query ✕

Upload Data

Hide Sections

Gemini

- GEMINI query** Querying the GEMINI database
- GEMINI set_somatic** Tag somatic mutations in a GEMINI database
- GEMINI fusions** Identify somatic fusion genes from a GEMINI database
- GEMINI amend** Amend an already loaded GEMINI database.
- GEMINI load** Loading a VCF file into GEMINI
- GEMINI annotate** the variants in an existing GEMINI database with additional information
- GEMINI database info** Retrieve information about tables, columns and annotation data stored in a GEMINI database
- GEMINI stats** Compute useful variant statistics
- GEMINI actionable_mutations** Retrieve genes with actionable somatic mutations via COSMIC and DGIdb

database. E.g. alt='G' or impact_severity = 'HIGH'.

Output format options

Type of report to generate

tabular (GEMINI default)

Add a header of column names to the output

Yes

(--header)

Set of columns to include in the variant report table

Custom (report user-specified columns)

Choose columns to include in the report

Select/Unselect all

- gene
- chrom
- start
- end
- ref
- alt
- impact
- impact_severity
- alternative allele frequency (max_aaf_all)

(--columns)

Additional columns (comma-separated)

type, gt_alt_freqs.TUMOR, gt_alt_freqs.NORMAL, ifnull(nullif(round(max_aaf_all,2),-1.0),0) AS MAF

Column must be specified by the exact name they have in the GEMINI database, e.g., is_exonic or num_hom_alt, but, for genotype columns, GEMINI wildcard syntax is supported. The order of columns in the list is maintained in the output.

Request drug-gene interaction info from DGIdb

No

```
type,  
gt_alt_freqs.TUMOR,  
gt_alt_freqs.NORMAL,  
ifnull(nullif(round(max_aaf_all,2),-1.0),0)  
AS MAF,  
gene,  
impact_so,  
aa_change,  
ifnull(round(cadd_scaled,2),'.') AS  
cadd_scaled,  
round(gerp_bp_score,2) AS gerp_bp,  
ifnull(round(gerp_element_pval,2),'.') AS  
gerp_element_pval,  
ifnull(round(hs_qvalue,2), '.') AS  
hs_qvalue,  
in_omim,  
ifnull(clinvar_sig, '.') AS clinvar_sig,  
ifnull(clinvar_disease_name, '.') AS  
clinvar_disease_name,  
ifnull(rs_ids, '.') AS dbsnp_ids,  
rs_ss,  
ifnull(cosmic_ids, '.') AS cosmic_ids,  
ifnull(overlapping_civic_url, '.') AS  
overlapping_civic_url,  
in_cgldb
```

c. Generating reports of genes affected by variants

Turning query results into gene-centered reports

GEMINI query Querying the GEMINI database (Galaxy Version 0.20.1+galaxy1) Favorite Versions Options

GEMINI database
96: GEMINI annotate CGI infos on data 54 and data 95

Only files with version 0.20.1 are accepted.

Build GEMINI query using
Advanced query constructor

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
v.somatic_status = 2 AND v.somatic_p <= 0.05 AND v.filter IS NULL GROUP BY g.gene
```

Formulate your query using SQL syntax. (-q)

Genotype filter expression

1: Genotype filter expression

Restrictions to apply to genotype values

```
gt_alt_freqs.NORMAL <= 0.05 AND gt_alt_freqs.TUMOR >= 0.10
```

(--gt-filter)

+ Insert Genotype filter expression

Sample filter expression

+ Insert Sample filter expression

Output format options

```
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
```

```
GROUP BY g.gene
```


d. Adding additional annotations to the gene-centered report

Adding UniProt cancer genes information

 Join two files (Galaxy Version 1.1.2)

1st file

   59: GEMINI query on data 55   

Column to use from 1st file

Column: 1 

2nd File

   43: Uniprot_Cancer_Genes.13Feb2019.txt   

Column to use from 2nd file

Column: 1 

Output lines appearing in

Both 1st & 2nd file, plus unpairable lines from 1st file. (-a 1) 

First line is a header line

Yes

Use if first line contains column headers. It will not be sorted.

Ignore case

No

Sort and Join key column values regardless of upper/lower case letters.

Value to put in unpaired (empty) fields


0

Email notification






Send an email notification when the job completes.

 Execute

Adding CGI biomarkers information

 **Join two files** (Galaxy Version 1.1.2) ★ 🔄 ▾






1st file

   60: Join on data 43 and data 59 ▾  

Column to use from 1st file

Column: 1 ▾

2nd File

   44: cgi_genes.txt ▾  

Column to use from 2nd file

Column: 1 ▾

Output lines appearing in

Both 1st & 2nd file, plus unpairable lines from 1st file. (-a 1) ▾

First line is a header line

Yes Use if first line contains column headers. It will not be sorted.

Ignore case


No Sort and Join key column values regardless of upper/lower case letters.

Value to put in unpaired (empty) fields

0

Email notification

Send an email notification when the job completes.

 **Execute**

Adding gene information from CIViC

 Join two files (Galaxy Version 1.1.2)

1st file



61: Join on data 44 and data 60



Column to use from 1st file

Column: 1

2nd File



45: 01-Feb-2019-GeneSummaries.tsv



Column to use from 2nd file

Column: 3

Output lines appearing in

Both 1st & 2nd file, plus unpairable lines from 1st file. (-a 1)

First line is a header line



Yes

Use if first line contains column headers. It will not be sorted.

Ignore case



No

Sort and Join key column values regardless of upper/lower case letters.

Value to put in unpaired (empty) fields

.

Email notification



Send an email notification when the job completes.

 Execute

Rearrange to get a fully annotated gene report

Column arrange by header name (Galaxy Version 0.2) ☆ Favorite & Versions ▼ Options

file to rearrange

110: Join on data 59 and data 109

Specify the first few columns by name

1: Specify the first few columns by name

column

gene

2: Specify the first few columns by name

column

chrom

3: Specify the first few columns by name

column

synonym

4: Specify the first few columns by name

column

hgnc_id

5: Specify the first few columns by name

column

entrez_id

6: Specify the first few columns by name

column

rvis_pct

7: Specify the first few columns by name

column

is_OG

8: Specify the first few columns by name

column

is_TS

9: Specify the first few columns by name

column

in_cgi_biomarkers

10: Specify the first few columns by name

column

clinvar_gene_phenotype

11: Specify the first few columns by name

column

gene_civic_url

12: Specify the first few columns by name

column

description

+ Insert Specify the first few columns by name