

Galaxy for virologist training

Exercise 6: Illumina Variant Calling 101

Title	Galaxy
Training dataset:	PRJEB43037 - In August 2020, an outbreak of West Nile Virus affected 71 people with meningoencephalitis in Andalusia and 6 more cases in Extremadura (south-west of Spain), causing a total of eight deaths. The virus belonged to the lineage 1 and was relatively similar to previous outbreaks occurred in the Mediterranean region. Here, we present a detailed analysis of the outbreak, including an extensive phylogenetic study. This is one of the outbreak samples.
Questions:	<ul style="list-style-type: none">• What is variant calling?• What is a vcf file?• How can I inspect a variant in a bam file to look for false positives?• How can I make a consensus genome based on a variant calling process?
Objectives:	<ul style="list-style-type: none">• Understand variant calling concept• Learn how to interpret a vcf file• Learn how to make a reference consensus genome.• Learn how to visualize mapping and variant calling results
Estimated time:	2h

Table of Contents

- [1. Description](#)
- [2. Upload data to galaxy](#)
 - [Training dataset](#)
 - [Create new history](#)
 - [Upload data](#)
- [3. Preprocess our reads.](#)
- [4. Map trimmed reads against the reference genome.](#)
- [5. Variant Calling.](#)
 - [Samtools mpileup](#)
 - [VarScan](#)
 - [VCF stats](#)
 - [Ivar variants](#)
 - [Lofreq](#)
 - [Insert indel qualities](#)
 - [Call variants](#)
- [Compare vcfs among callers](#)
 - [Visualize datasets.](#)
- [7. Consensus genome](#)
 - [Bcftools consensus](#)
 - [Ivar Consensus](#)

1. Description


After mapping, when we have a re-sequencing experiment, the next step usually comprises the variants calling step. Variant calling software tries to identify variants, positions that differ in our reads compared to a reference genome. We may want to have a consensus genome as well, which is obtained by including the variants we just identified in the published reference genome. We are going to address this type of analysis in this tutorial.

2. Upload data to galaxy

Training dataset

- Experiment info: PRJEB43037, WGS, Illumina MiSeq, paired-end
- Fastq R1: [ERR5310322_1](#) - url :
`ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fastq.gz`
- Fastq R2: [ERR5310322_2](#) url :
`ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_2.fastq.gz`
- Reference genome NC_009942.1: [fasta](#) -- [gff](#)

Create new history

- Click the  icon at the top of the history panel and create a new history with the name `variant calling 101 tutorial` as explained [here](#)

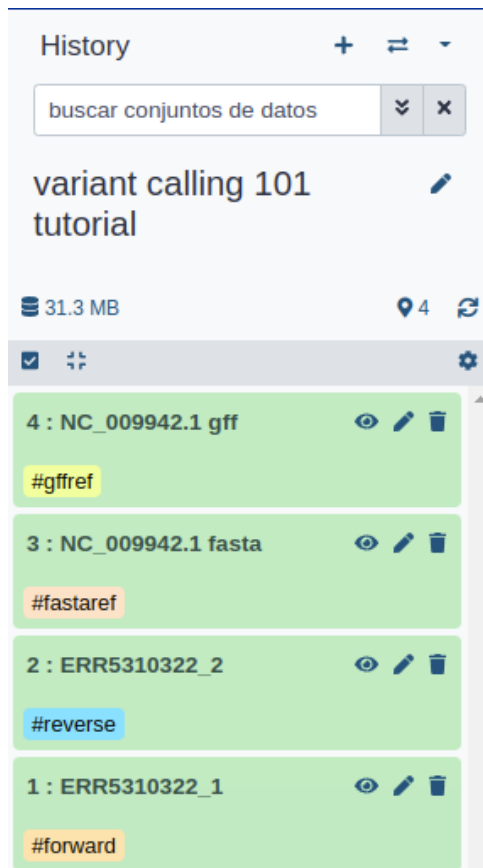
Upload data

Follow the same instructions [here](#)

```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000
```

Rename the data as follows:

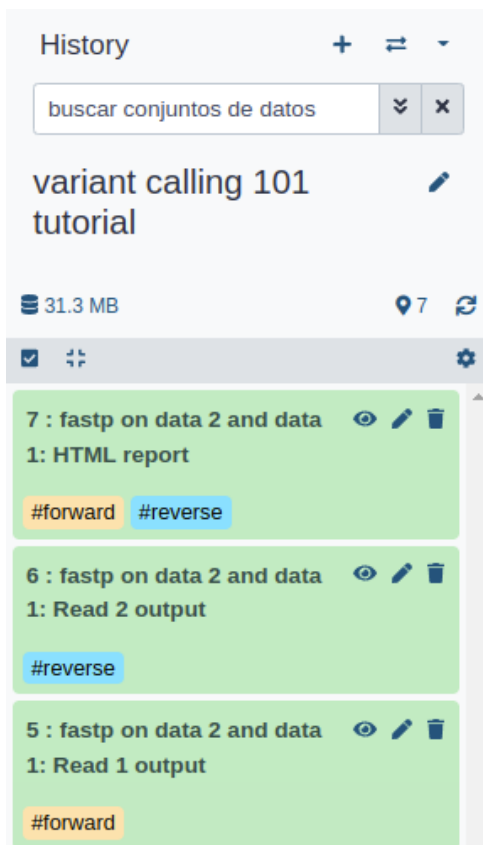
- ERR5310322_1.fastq.gz to ERR5310322_1 with tag #forward
- ERR5310322_2.fastq.gz to ERR5310322_2 with tag #reverse
- GCF_000875385.1_ViralProj30293_genomic.fna.gz to NC_009942.1 fasta with tag #fastaref
- GCF_000875385.1_ViralProj30293_genomic.gff.gz to NC_009942.1 gff with tag #gffref



3. Preprocess our reads.

Follow instructions [here](#)

Then, fix fastp tags on the output data to be as follows:



4. Map trimmed reads against the reference genome.

Follow: 1. Is this single or paired library: paired. 2. FASTA/Q file #1 : fastp Read 1 output #forward 3. FASTA/Q file #2 : fastp Read 2 output #reverse 4. Will you select a reference genome from your history or use a built-in index? : Use a genome from the history and build index. 5. Do you want to use presets? : Very sensitive local. This setting will hugely affect the mapping results, depending on the dataset/experiment must be tweaked (read [bowtie2 manual](#)) - Save the bowtie2 mapping statistics to the history

Galaxy Europe

Flujo de Trabajo Visualizar Datos Compartidos Ayuda Usuario

Using 0%

Herramientas

bowtie2 1

Cargar Datos

Show Sections

Bowtie2 - map reads against reference genome 2

SALSA scaffold long read assemblies with Hi-C

bamPEFragmentSize Estimate the predominant cDNA fragment length from paired-end sequenced BAM/CRAM files

Extract the marker sequences and metadata from the MetaPhlAn database

TB-Profiler Profile Infer strain types and drug resistance markers from sequences

HUMANN to profile presence/absence and abundance of microbial pathways and gene families

MetaPhlAn to profile the composition of microbial communities

MaxBin2 clusters metagenomic contigs into bins

InStrain Profile Creates an inStrain profile (microdiversity analysis) from a mapping file

hicBuildMatrix create a contact matrix

Man with miniman? & fact pairwise

Bowtie2 - map reads against reference genome (Galaxy Version 2.5.0+galaxy0)

Is this single or paired library

Paired-end 3

FASTA/Q file #1

5: fastp on data 2 and data 1: Read 1 output 4

Must be of datatype "fastqsanger" or "fasta"

FASTA/Q file #2

6: fastp on data 2 and data 1: Read 2 output 5

Must be of datatype "fastqsanger" or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Do you want to set paired-end options?

No

See "Alignment Options" section of Help below for information

Will you select a reference genome from your history or use a built-in index?

Use a genome from the history and build index

Built-ins were indexed using default options. See "Indexes" section of help below

Select reference genome

3: NC_009942.1 fasta (as fasta)

Set read groups information?

History

variant calling 101 tutorial

31.3 MB

7: fastp on data 2 and data 1: HTML report

#forward #reverse

6: fastp on data 2 and data 1: Read 2 output

#reverse

5: fastp on data 2 and data 1: Read 1 output

#forward

4: NC_009942.1 gff

#gffref

3: NC_009942.1 fasta

#fastaref

2: ERR5310322_2

#reverse

1: ERR5310322_1

Galaxy Europe

Flujo de Trabajo
Visualizar
Datos Compartidos
Ayuda
Usuario
Using 0%

Herramientas

bowtie2

Cargar Datos

Show Sections

Bowtie2 - map reads against reference genome

SALSA scaffold long read assemblies with Hi-C

bamPEFragmentSize Estimate the predominant cDNA fragment length from paired-end sequenced BAM/CRAM files

Extract the marker sequences and metadata from the MetaPhlAn database

TB-Profiler Profile Infer strain types and drug resistance markers from sequences

HUMANN to profile presence/absence and abundance of microbial pathways and gene families

MetaPhlAn to profile the composition of microbial communities

MaxBin2 clusters metagenomic contigs into bins

InStrain Profile Creates an inStrain profile (microdiversity analysis) from a mapping file

hicBuildMatrix create a contact matrix

Man with miniman? A fast pairwise

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

1: Default setting only

Do you want to use presets?

☐ No, just use defaults
☐ Very fast end-to-end (--very-fast)
☐ Fast end-to-end (--fast)
☐ Sensitive end-to-end (--sensitive)
☐ Very sensitive end-to-end (--very-sensitive)
☐ Very fast local (--very-fast-local)
☐ Fast local (--fast-local)
☐ Sensitive local (--sensitive-local)
☒ Very sensitive local (--very-sensitive-local)

Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets.

Do you want to tweak SAM/BAM Options?

No

See "Output Options" section of Help below for information

Save the bowtie2 mapping statistics to the history

☒ Yes

Email notification

☐ No

Send an email notification when the job completes.

Execute

Bowtie2 Overview

Bowtie2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly

History

buscar conjuntos de datos

variant calling 101 tutorial

31.3 MB

7

7 : fastp on data 2 and data 1: HTML report

#forward #reverse

6 : fastp on data 2 and data 1: Read 2 output

#reverse

5 : fastp on data 2 and data 1: Read 1 output

#forward

4 : NC_009942.1 gff

#gffref

3 : NC_009942.1 fasta

#fastaref

2 : ERR5310322_2

#reverse

1 : ERR5310322_1

5. Variant Calling.

Samtools mpileup

1. Search samtools mpileup in the search toolbox, scroll down and select `Samtools mpileup multi-way pileup of variants`
2. Bam files: Bowtie2 bam file
3. Use reference: Use reference/genome from history. NC_009942.1.
4. Set advanced options: Advanced
5. Disable read-pair overlap detection: Yes
6. Disable BAQ (per-Base Alignment Quality), see below: Yes
7. Do not discard anomalous read pairs: Yes
8. max per-file depth; avoids excessive memory usage: 0
9. Minimum base quality for a base to be considered: 20
10. Click execute and wait.

Samtools mpileup multi-way pileup of variants

(Galaxy Version 2.1.5)

Favorite

Versions

Options

BAM file(s)

7: Bowtie2 on data 3, data 5, and data 4: alignments

7: Bowtie2 on data 3, data 5, and data 4: alignments

7: Bowtie2 on data 3, data 5, and data 4: alignments

7: Bowtie2 on data 3, data 5, and data 4: alignments

Use a reference sequence

Use a genome/index from the history

Reference

3: NC_009942.1

3: NC_009942.1

3: NC_009942.1

(-f)

Disable read-pair overlap detection

☒ Yes

(-x/--ignore-overlaps)

Do not discard anomalous read pairs

☒ Yes

(-A/--count-orphans)

Disable BAQ (per-Base Alignment Quality), see below

☒ Yes

(-B/--no-BAQ)

Minimum base quality for a base to be considered

(-Q/--min-BQ)

1. Click the :eye: icon on the history and inspect the mpileup output.

VarScan

1. Search `VarScan Mpileup` in the search toolbox.
2. Samtools pileup dataset: samtools mpileup output
3. Minimum read depth: 10
4. Minimum supporting reads: 5
5. Minimum base quality at a position to count a read: 20
6. Minimum variant allele frequency threshold: 0,75
7. Default p-value threshold for calling variants: 0,05
8. Click execute and wait

VarScan mpileup for variant detection (Galaxy
Version 2.4.3.1)

☆ Favorite

🔗 Versions

▼ Options

Samtools pileup dataset



16: samtools mpileup on data 3 and data 7 pileup



Analysis type

single nucleotide variation

Minimum coverage



Minimum depth at a position to make a call (--min-coverage)

Minimum supporting reads



Minimum number (default: 2) of variant-supporting reads at a position required to make a call (--min-reads2)

Minimum base quality



Minimum variant allele frequency (default: 0.01) required for calling a variant (--min-var-freq)

0.75

Minimum variant allele frequency (default: 0.75) required for calling a homozygous genotype (--min-freq-for-hom)

0.05

```
(--p-value)
```

1. Click the `:eye:` icon and inspect the vcf file.

VCF stats

1. Search `bcftools stats` in the search toolbox.
2. VCF/BCF Data: varscan vcf output.
3. Click execute and wait.
4. Click the `:eye:` icon and inspect the stats.

- ▶ How many variants do we have in our vcf file?

Ivar variants

1. Search `ivar variants` in the search toolbox.
2. Samtools pileup dataset: samtools mpileup output.
3. Bam file: bowtie bam output
4. Reference: NC_009942.1
5. Minimum quality score threshold to count base: 20
6. Minimum frequency threshold: 0.75
7. Output format: Both tabular and vcf
8. Click execute and wait.

Galaxy Europe

Flujo de Trabajo

Visualizar

Datos Compartidos

Ayuda

Usuario

Herramientas

ivar

1

x

Cargar Datos

Show Sections

ivar removereads

Remove reads from trimmed BAM file

ivar variants

Call variants from aligned BAM file

2

ivar consensus

Call consensus from aligned BAM file

ivar getmasked

Detect primer mismatches and get primer indices for the amplicon to be masked

ivar filtervariants

Filter variants across replicates or multiple samples aligned using the same reference

ivar trim

Trim reads in aligned BAM

Freyja: Call variants

and get sequencing depth information

Freyja: Bootstrapping

method

Freyja: Aggregate and visualize

demixed results

Freyja: Demix

lineage abundances

Multivariate

PCA, PLS and OPLS

Univariate

Univariate statistics

FLUJOS DE TRABAJO

Todos los flujos de trabajo

ivar variants

Call variants from aligned BAM file (Galaxy Version 1.3.1+galaxy2)

Bam file

8: Bowtie2 on data 3, data 6, and data 5: alignments

Aligned reads, to trim primers and quality

Reference

3: NC_009942.1 fasta (as fasta)

Minimum quality score threshold to count base

20

(-q)

Minimum frequency threshold

0,75

(-t)

Output format

Both Tabular and VCF

In VCF only output variants that PASS all filters

Yes

(--pass_only)

Email notification

No

Send an email notification when the job completes.

Execute

ivar uses the output of the samtools mpileup command to call variants - single nucleotide variants(SNVs) and indels. In order to call variants correctly, the reference file used for alignment must be passed to iVar using the -r flag. The output of samtools pileup is piped into iVar variants to generate a .tsv file with the variants. There are two parameters that can be set for variant calling using iVar - minimum quality(Default: 20) and minimum frequency(Default: 0.03). Minimum quality is the minimum quality

Lofreq

Insert indel qualities

1. Search `Insert indel qualities` in the search toolbox. Select Insert indel qualities with lofreq.
2. Reads: bowtie2 bam output.
3. Click execute and wait.

Insert indel qualities into a BAM file (Galaxy Version 2.1.5+galaxy0)

☆ Favorite🔗 Versions▼ Options

Reads

7: Bowtie2 on data 3, data 5, and data 4: alignments▼

Indel calculation approach

Uniform▼

Indel quality to add

30

Should probably not be left at the default value

Separate deletion quality

Leave blank to use the same values for insertions and deletions

Call variants

1. Search `Lofreq` in the search toolbox. Select Call variants with lofreq.
2. Input reads in BAM format: indel qualities bam output.
3. Choose the source for the reference genome: History. NC_009942.1
4. Types of variants to call: SNVs and INDELS
5. Variant calling parameters: Configure settings
6. Minimal coverage: 10
7. Minimum baseQ: 20
8. Minimum baseQ for alternate bases: 20
9. Click execute and wait.

Call variants with LoFreq (Galaxy Version 2.1.5+galaxy1)

☆ Favorite🔗 Versions▼ Options

Input reads in BAM format

7: Bowtie2 on data 3, data 5, and data 4: alignments▼

Choose the source for the reference genome

History▼

Reference

3: NC_009942.1 (as fasta)▼

Reference sequence (--ref)

Call variants across

Whole reference▼

Types of variants to call

SNVs and indels▼

Coverage

Minimal coverage


10

Do not attempt variant calling at sites that are not covered by at least this number of reads (default: 1) (--min-cov)

Coverage cap

1000000

For efficiency, don not consider more than this number of reads at any site (default: 1,000,000) (--max-depth)

Base-calling quality 

Minimum baseQ

For variant calling at any given site, do not consider reads for which the base at that site has a base quality less than this value (default: 6) (--min-bq)

Minimum baseQ for alternate bases

For variant calling at any given site, do not consider reads that support a non-reference allele at the site if that base has a base quality less than this value (default: 6). Note: this setting will have no effect if the specified value is less than the general Minimum baseQ above. (--min-alt-bq)

Base quality to use for alternate bases

Compare vcfs among callers

Visualize datasets.

1. Search `upSet diagram` in the search toolbox.
2. Select input files for which to produce intersections: select vcf from varscan, vcf from lofreq filter and vcf from ivar variants.
3. Click execute and wait.
4. Click the `:eye:` icon and check the diagram.




► How many variants differ among the vcfs?

7. Consensus genome



Bcftools consensus



1. Search `bcftools consensus` in the search toolbox.
2. VCF/BCF Data: varscan vcf output.
3. Choose a reference genome: use genome/reference from history. Select NC_009942.1.
4. Click execute and wait.

bcftools consensus Create consensus sequence by applying VCF variants to a reference fasta file (Galaxy Version 1.9+galaxy1)

 Favorite
  Versions
  Options



VCF/BCF Data






Choose the source for the reference genome

Reference genome

Note: for this example we are not going to mask any position with low coverage, this will be addressed in the exercise 8, with a real example.

Ivar Consensus

1. Search `ivar consensus` in the search toolbox.
2. Bam file: bowtie bam output.
3. Use N instead of - for regions with less than minimum coverage: Yes




ivar consensus Call consensus from aligned BAM file (Galaxy Version 1.3.1+galaxy0)

☆ Favorite



🔄 Versions

▼ Options

Bam file



7: Bowtie2 on data 3, data 5, and data 4: alignments ▼



Aligned reads, to trim primers and quality

Minimum quality score threshold to count base

(-q)

Minimum frequency threshold

0 - Majority or most common base

0.2 - Bases that make up atleast 20% of the depth at a position

0.5 - Strict or bases that make up atleast 50% of the depth at a position

0.9 - Strict or bases that make up atleast 90% of the depth at a position

1 - Identical or bases that make up 100% of the depth at a position. Will have highest ambiguities (-t)

Minimum depth to call consensus

Here is the galaxy history for this exercise:

<https://usegalaxy.eu/u/smonzon/h/variant-calling-101-tutorial-1>