

# Galaxy for virologist training Exercise 7: Illumina Variant Annotation 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"><li>• Which effects have variants in the genome?</li></ul>
Objectives:	<ul style="list-style-type: none"><li>• Understand the importance of variants effect significance.</li></ul>
Estimated time:	1h

## 1. Description

After performing variant calling, we want to know which is the importance of the variants in the viral genome. In order to give sense to the variants, we need to know in which gene they are, and which are their effects.

## 2. Upload data to galaxy

### Training dataset

- Experiment info: PRJEB43037, WGS, Illumina MiSeq, paired-end
- Fastq R1: [ERR5310322\\_1](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fastq.gz) - url :  
[ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322\\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fastq.gz)
- Fastq R2: [ERR5310322\\_2](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_2.fastq.gz) url :  
[ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322\\_2.fastq.gz](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 [mapping 101 tutorial](#) as explained [here](#)

### Upload data

Follow the same instructions [here](#)

```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_2.fastq.gz
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_ViralProj30293/GCF_000875385.1_ViralProj30293_genomic.fna.gz
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_ViralProj30293/GCF_000875385.1_ViralProj30293_genomic.gff.gz
```

### 3. Preprocess our reads

Follow instructions [here](#)

### 4. Map our reads against our reference genome

Follow instructions [here](#)

### 5. Variant Calling

Follow instructions [here](#)

### 6. Variants annotation

Snpeff build

1. Search **snpeff build** in the search toolbox.
2. Name of the database: WestNile.
3. Input annotations are in: GFF
4. GFF dataset to build database from: NC\_009942.1 gff
5. Choose the source of the reference genome: History. NC\_009942.1 fasta.
6. Click execute and wait.

**Name for the database**

WestNile

For E. coli K12 you may want to use 'Eck12' etc.

**Input annotations are in**

☐ GenBank

☒ GFF

☐ GTF

Specify format for annotations you are using to create SnpEff database

**GFF dataset to build database from**

43: NC\_009942.1 (as gff3)

This GFF file will be used to generate snpEff database

**Choose the source for the reference genome**

History

**Genome in FASTA format**

3: NC\_009942.1

This dataset is required for generating SnpEff database. See help section below.

Snpeff eff

- 1. Search **snpeff** **eff** in the search toolbox.
- 2. Sequence changes (SNPs, MNPs, InDels): ivar vcf file
- 3. Create CSV report, useful for downstream analysis (-csvStats): Yes.
- 4. Genome source: Custom snpEff database in your history. Snpeff build output.
- 5. Click execute and wait.

**Sequence changes (SNPs, MNPs, InDels)**

34: ivar variants VCF on data 3 and data 7

**Input format**

VCF

**Output format**

VCF (only if input is VCF)

**Create CSV report, useful for downstream analysis (-csvStats)**

☒

Yes

**Genome source**

Custom snpEff database in your history

**SnpEff4.3 Genome Data**

44: SnpEff4.3 database for WestNile

**Select genetic code for this sequence**

Standard

- 6. Click the **:eye:** icon in the SnpEff html output and check the results.

3 / 4

SnpSift: transfrom vcf snpeff to table.

- 1. Search **SnpSift ExtractFields** in the search toolbox.
- 2. Variant input file in VCF format: snpeff eff vcf output.
- 3. Fields to extract: **CHROM POS ID REF ALT FILTER ANN[\*].EFFECT ANN[\*].GENE ANN[\*].FEATURE ANN[\*].HGVS\_C ANN[\*].HGVS\_P**
- 4. One effect per line: Yes.
- 5. Click execute and wait.
- 6. Click the :eye: icon in the snpsift output and check the results.

**SnpSift Extract Fields** from a VCF file into a tabular file (Galaxy Version 4.3+t.galaxy0)

☆ Favorite

Versions

▼ Options

Variant input file in VCF format

45: SnpEff eff: on data 44 and data 34

↕

Fields to extract

CHROM POS ID REF ALT FILTER ANN[\*].EFFECT ANN[\*].GENE ANN[\*].FEATURE ANN[\*].HGVS\_C ANN[\*].HGVS\_P

Separated by spaces. See help below for an explanation

One effect per line

Yes

When variants have more than one effect, lists one effect per line, while all other parameters in the line are repeated across mutiple lines

multiple field separator

Separate multiple fields in one column with this character, e.g. a comma, rather than a column for each of the multiple values (-s)

empty field text

Galaxy history for this exercise: <https://usegalaxy.eu/u/smonzon/h/variant-calling-101-tutorial>