1.0 Introduction

SnpEff is a variant annotation and effect prediction tool. It annotates variants based on their genomic locations and predicts coding effects. Annotated genomic locations include intronic, untranslated region, upstream, downstream, splice site, or intergenic regions. Coding effects such as synonymous or non-synonymous amino acid replacement, start codon gains or losses, stop codon gains or losses, or frame shifts can be predicted. SnpEff is open source, platform independent and freely available for all users. The main features of SnpEff include:

- Supports over 38,000 genomes.
- Standard ANN annotation format
- Cancer variants analysis
- GATK compatible (-o gatk)
- HGVS notation
- Sequence Ontology standardized terms

In addition to SnpEff, there are other recently developed programs for annotating genomic variants, most notably “Annotate Variation” (ANNOVAR) and “Variant Annotation, and Analysis and Search Tool” (VAAST). However, SnpEff differs from these programs in that it is an open source for all users, permits annotation of more genome versions, natively supports Variant Call Format (VCF) files and it is marginally faster (although the speeds of SnpEff, ANNOVAR, and VAAST are comparable).

2.0 How to install

SnpEff requires that you have Java v1.8 or later installed (any modern operating system has it). The amount of memory used can vary significantly depending on genome size and data analysis type you are doing. For large genomes, such as the human genome, you'll probably need at least 4Gb of memory. Installing SnpEff is very easy, you just have to uncompress the ZIP file.

```
# Go to home dir
cd

# Download latest version
wget https://snpeff.blob.core.windows.net/versions/snpEff_latest_core.zip

# Unzip file
unzip snpEff_latest_core.zip
```

By default, SnpEff automatically downloads and installs the database for you, so you don't need to do it manually. For more information, please refer to https://pcingola.github.io/SnpEff/download/

3.0 How to use SnpEff

3.1 Data input

Three input formats supported by SnpEff are variant call format (VCF), tab separated TXT format; and the SAMtools Pileup format. VCF was created by the 1000 Genomes project, and it is currently the de facto standard for variants in sequencing applications. The TXT and Pileup formats are currently deprecated and being phased out.

Usage:

```
java -Xmx8g -jar snpEff.jar GRCh37.75 examples/test.chr22.vcf > test.chr22.ann.vcf
```

By default, the amount of memory set by a java process is set too low. If you don't assign more memory to the process, you will most likely have an "OutOfMemory" error. You should set the amount of memory in your java virtual machine to, at least, 2 Gb. This can be easily done using the Java command line option -Xmx. Example of using 4Gb:

```
java -Xmx8g snpEff.jar hg19 path/to/your/files/snps.vcf
```

One of the first thing SnpEff must do is to load the database. Usually, it takes from a few seconds to a couple of minutes, depending on database size. Complex databases, like human, require more time to load. After the database is loaded, SnpEff can analyse thousands of variants per second.

3.3 Output file interpretation

SnpEff also supports two output formats, TXT and VCF. The output information provided in both formats includes three main groups:

I. Variant information (genomic position, the reference and variant sequences, change type, heterozygosity, quality and coverage)
II. Genetic information (gene Id, gene name, gene biotype, transcript ID, exon ID, exon rank)
III. Effect information (effect type, amino acid changes, codon changes, codon number in CDS, codon degeneracy, etc.)

Whenever multiple transcripts for a gene exist, the effect and annotations on each transcript are reported, so one variant can have multiple output lines. When using the VCF format, the functional annotations information is added to the INFO field using an ANN tag. VCF INFO field name ANN, stands for 'annotations'. Here is a description of the meaning of each sub-field:

I. **Allele (or ALT):** In case of multiple ALT fields, this helps to identify which ALT we are referring to.
II. **Annotation (a.k.a. effect):** Annotated using Sequence Ontology terms. Multiple effects can be concatenated using '&'.
III. **Putative_impact**: A simple estimation of putative impact / deleteriousness : {HIGH, MODERATE, LOW, MODIFIER}
IV. **Gene Name**: Common gene name (HGNC). Optional: use closest gene when the variant is "intergenic".
V. **Gene ID**: Gene ID
VI. **Feature type:** Which type of feature is in the next field (e.g. transcript, motif, miRNA, etc.). It is preferred to use Sequence Ontology (SO) terms, but 'custom' (user defined) are allowed.
VII. **Feature ID**: Depending on the annotation, this may be: Transcript ID (preferably using version number), Motif ID, miRNA, ChipSeq peak, Histone mark, etc. Note: Some features may not have ID (e.g. histone marks from custom Chip-Seq experiments may not have a unique ID).
VIII. **Transcript biotype:** The bare minimum is at least a description on whether the transcript is {"Coding", "Noncoding"}. Whenever possible, use ENSEMBL biotypes.
IX. **Rank / total:** Exon or Intron rank / total number of exons or introns.
X. **HGVS.c:** Variant using HGVS notation (DNA level)

XI. **HGVS.p**: If variant is coding, this field describes the variant using HGVS notation (Protein level). Since transcript ID is already mentioned in 'feature ID', it may be omitted here.

XII. **cDNA_position / cDNA_len:** Position in cDNA and trancript's cDNA length (one based).

XIII. **CDS_position / CDS_len:** Position and number of coding bases (one based includes START and STOP codons).

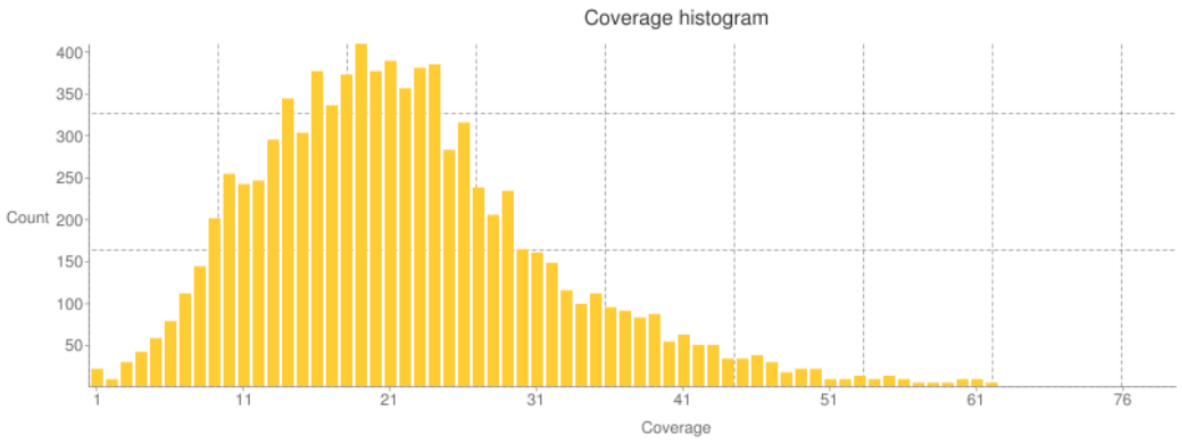
XIV. **Protein_position / Protein_len:** Position and number of AA (one based, including START, but not STOP).

XV. **Distance to feature:** All items in this field are options, so the field could be empty.

- Up/Downstream: Distance to first / last codon
- Intergenic: Distance to closest gene
- Distance to closest Intron boundary in exon (+/- up/downstream). If same, use positive number.
- Distance to closest exon boundary in Intron (+/- up/downstream)
- Distance to first base in MOTIF
- Distance to first base in miRNA
- Distance to exon-intron boundary in splice_site or splice _region
- ChipSeq peak: Distance to summit (or peak center)
- Histone mark / Histone state: Distance to summit (or peak center)

XVI. **Errors, Warnings or Information messages:** Add errors, warnings or informative message that can affect annotation accuracy. It can be added using either 'codes' (as shown in column 1, e.g. W1) or 'message types' (as shown in column 2, e.g. WARNING_REF_DOES_NOT_MATCH_GENOME). All these errors, warnings or information messages messages are optional.

SnpEff creates an additional output file showing overall statistics. The program performs some statistics and saves them to the file 'snpEff_summary.html' on the directory where snpEff is being executed. You can see the file, by opening it in your browser. In the stats file, you can see coverage histogram plots like this one:



"Effects by type" vs "Effects by region"

SnpEff annotates variants. Variants produce effect of difference "types" (e.g. NON_SYNONYMOUS_CODING, STOP_GAINED). These variants affect regions of the genome (e.g. EXON, INTRON). The two tables count how many effects for each type and for each region exists.

E.g.: In an EXON region, you can have all the following effect types: NON_SYNONYMOUS_CODING, SYNONYMOUS_CODING, FRAME_SHIFT, STOP_GAINED, etc.

The complicated part is that some effect types affect a region that has the same name (yes, I know, this is confusing).

E.g.: In a UTR_5_PRIME region you can have UTR_5_PRIME and START_GAINED effect type.

This means that the number of both tables are not exactly the same, because the labels don't mean the same. See the next figure as an example:

Type

Type (alphabetical order)	Count	Percent
DOWNSTREAM	2,093	1.766%
INTERGENIC	26,314	22.204%
INTRAGENIC	78	0.066%
INTRON	54,238	45.767%
NON_SYNONYMOUS_CODING	237	0.2%
NON_SYNONYMOUS_START	1	0.001%
SPLICE_SITE_DONOR	4	0.003%
START_GAINED	57	0.048%
STOP_GAINED	3	0.003%
STOP_LOST	1	0.001%
SYNONYMOUS_CODING	378	0.319%
TRANSCRIPT	32,163	27.14%
UPSTREAM	2,102	1.774%
UTR_3_PRIME	690	0.582%
UTR_5_PRIME	149	0.126%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	2,093	1.766%
EXON	620	0.523%
INTERGENIC	26,314	22.204%
INTRON	54,238	45.767%
NONE	32,241	27.206%
SPLICE_SITE_DONOR	4	0.003%
UPSTREAM	2,102	1.774%
UTR_3_PRIME	690	0.582%
UTR_5_PRIME	206	0.174%

So the number of effects that affect a UTR_5_PRIME region is 206. Of those, 57 are effects type START_GAINED and 149 are effects type UTR_5_PRIME. How exactly are effect type and effect region related? See the following table:

Effect Type	Region
NONE CHROMOSOME CUSTOM CDS	NONE
INTERGENIC INTERGENIC_CONSERVED	INTERGENIC
UPSTREAM	UPSTREAM
UTR_5_PRIME UTR_5_DELETED START_GAINED	UTR_5_PRIME
SPLICE_SITE_ACCEPTOR	SPLICE_SITE_ACCEPTOR

SPLICE_SITE_DONOR	SPLICE_SITE_DONOR
SPLICE_SITE_REGION	SPLICE_SITE_REGION
INTRAGENIC START_LOST SYNONYMOUS_START NON_SYNONYMOUS_START GENE TRANSCRIPT	EXON or NONE
EXON EXON_DELETED NON_SYNONYMOUS_CODING SYNONYMOUS_CODING FRAME_SHIFT CODON_CHANGE CODON_INSERTION CODON_CHANGE_PLUS_CODON_INSERTION CODON_DELETION CODON_CHANGE_PLUS_CODON_DELETION STOP_GAINED SYNONYMOUS_STOP STOP_LOST RARE_AMINO_ACID	EXON
INTRON INTRON_CONSERVED	INTRON
UTR_3_PRIME UTR_3_DELETED	UTR_3_PRIME
DOWNSTREAM	DOWNSTREAM
REGULATION	REGULATION

SnpEff also generates a TXT (tab separated) file having counts of number of variants affecting each transcript and gene. By default, the file name is snpEff_genes.txt, but it can be changed using the -stats command line option. Here is an example of this file:

```
$ head snpEff_genes.txt
# The following table is formatted as tab separated values.
#GeneName   GeneId  TranscriptId    BioType variants_impact_HIGH    variants_impact
AC000029.1  ENSG00000221069 ENST00000408142 miRNA   0   0   0   2   0   0   0   2
AC000068.5  ENSG00000185065 ENST00000431090 antisense   0   0   0   1   0   0   0
AC000081.2  ENSG00000230194 ENST00000433141 processed_pseudogene    0   0   0   8
AC000089.3  ENSG00000235776 ENST00000424559 processed_pseudogene    0   0   0   1
AC002472.1  ENSG00000269103 ENST00000547793 protein_coding  0   0   0   6   0   0
AC002472.11 ENSG00000226872 ENST00000450652 antisense   0   0   0   13  0   0   0
AC002472.13 ENSG00000187905 ENST00000342608 protein_coding  0   1   6   1   0   0
AC002472.13 ENSG00000187905 ENST00000442047 protein_coding  0   1   6   1   0   0
```

The columns in this table are:

Column name	Meaning
GeneName	Gene name (usually HUGO)
GeneId	Gene's ID
TranscriptId	Transcript's ID
BioType	Transcript's bio-type (if available)
	The following column is repeated for each impact {HIGH, MODERATE, LOW, MODIFIER}
variants_impact_*	Count number of variants for each impact category
	The following column is repeated for each annotated effect (e.g. missense_variant, synonymous_variant, stop_lost, etc.)
variants_effect_*	Count number of variants for each effect type
	The following columns are repeated for several genomic regions (DOWNSTREAM, EXON, INTRON, UPSTREAM, etc.)
bases_affected_*	Number of bases that variants overlap genomic region
total_score_*	Sum of scores overlapping this genomic region. Note: Scores are only available when input files are type 'BED' (e.g. when annotating ChipSeq experiments)
length_*	Genomic region length

-End-