VarAnnot v1.0

**A Python Variant Annotator for VCF v4.0 files from Platypus v0.8.1 Variant Caller**

**Introduction**

This is a simple VCF annotator to parse VCF v4.0 files generated by Platypus v0.8.1 and annotate human genetic variants according to a patient’s genotype.

This tool adds information on consequence of the variant on protein sequence, rsID, depth of coverage at the site of variation, depth of ALT allele at the site of variation, allele balance ratio, EXAC frequency of the variant if available, associated gene and transcripts.

The data are fetched from Exome Aggregation Consortium (EXAC) and ClinVar databases. This piece of software is inspired from Ensemble’s Variant Effect Predictor (VEP) and Varant tools.

**Requirements**

This software was coded for **python v2.7.10**. This program requires the following python libraries:

* Operator
* Optparse
* Collections
* Requests v2.11.1 -- HTTP library
* PyVCF v0.6.8 -- VCF parser
* Pysam v0.9.1.4 – a Samtools wrapper to query VCF files
* Clinvar.vcf.gz & Clinvar.vcf.gz.tbi (Included in the bin folder)

The packages can be installed with the commands:

pip install pyvcf **(or)** sudo pip install pyvcf

pip install requests **(or)** sudo pip install requests

pip install pysam **(or)** sudo pip install pysam (**Works only for Linux/Mac OS**)

**Usage**

python /var\_annot\_v1.0/ varAnnot.py -h

**Input/Output**

Input – path to v4.0 VCF from Platypus v0.8.1

Output – samplename\_annotated.txt in the same path as the input folder unless specified

**Features:**

* Annotates only variants present in given sample by genotype
* Code can be modified to handle multiple samples
* Features obtained from EXAC can be customized
* Handles EXAC annotations from multiple genes with same consequence type for a variant
* Handles non-chromosome scaffolds

**Challenges:**

* Rank annotations by subtype of Impact class
* Platform compatibility of python libraries used (Pysam)
* Customize

**Annotations:**

1. **Consequence of the Variant on Protein Sequence (Adapted from VEP)**

For each variant, the ensemble transcript is identified and then using Sequence Ontology (SO) terms, the effects that the allele may have on the transcript is obtained. For variants that may have more than one predicted impact, the impact that might be more severe is selected. If the variant affects more than one gene, both the genes are provided in the annotation. Consequence classes in the order of their severity are shown in the table below:

**Consequence Classes of Variants on Protein Function**

|  |  |  |
| --- | --- | --- |
| Class Rank | Impact Class | Impact on Protein |
| 1 | High  (The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay.) | Transcript\_ablation  Splice\_acceptor\_variant  Splice\_donor\_variant  Stop\_gained  Frameshift\_variant  Stop\_lost  Start\_lost  Transcript\_amplification |
| 2 | Moderate  (A non-disruptive variant that might change protein effectiveness.) | Inframe\_insertion  Inframe\_deletion  Missense\_variant  Protein\_altering\_variant |
| 3 | Low  (Assumed to be mostly harmless or unlikely to change protein behaviour.) | Splice\_region\_variant  Incomplete\_terminal\_codon\_variant  Stop\_retained\_variant  Synonymous\_variant |
| 4 | Modifier  (Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact.) | Coding\_sequence\_variant  Mature\_miRNA\_variant  5\_prime\_UTR\_variant  3\_prime\_UTR\_variant  Non\_coding\_transcript\_exon\_variant  Intron\_variant  NMD\_transcript\_variant  Non\_coding\_transcript\_variant  Upstream\_gene\_variant  Downstream\_gene\_variant  TFBS\_ablation  TFBS\_amplification  TF\_binding\_site\_variant  Regulatory\_region\_ablation  Regulatory\_region\_amplification  Feature\_elongation  Regulatory\_region\_variant  Feature\_truncation  Intergenic\_variant |

1. **Depth of Coverage at the Site of Variation**

This is the total number of reads at the site of variation, obtained from the attribute:

##FORMAT=<ID=NR> attribute from the input VCF.

1. **Allele Depth**

This is the number of reads at the site of variation that support the variant, obtained from the attribute: ##FORMAT=<ID=NV> attribute from the input VCF.

1. **Allele Balance Percent**

This is defined as the proportion of reads that are supporting the variant vs. the reference:

Allele Balance (AB) = (# ALT reads) \*100/ (# REF + ALT reads)

= (NV\*100)/ (NR)

**Contact**

For any queries regarding the execution of the script please contact: Gayathri Rajan (gayathri.rajan89@gmail.com).